2019

Bayesian hierarchical modeling for disease outbreaks

Nehemias Ulloa
Iowa State University

Bayesian hierarchical modeling for disease outbreaks

by

Nehemias Ulloa

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Jarad Niemi, Major Professor
Alicia L. Carriquiry
Daniel J. Nordman
Chong Wang
Emily J. Berg

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation/thesis. The Graduate College will ensure this dissertation/thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2019

DEDICATION

This dissertation is dedicated to my wife Breanne and to my family. Without their support and prayers, I would not have been able to complete this work. I would especially like to thank my mom, Claudia Ulloa, for leading by example to push myself in my education and my dad, Miguel Ulloa, for showing me it is never to late to keep learning.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

This thesis would not have been possible without my advisor, Dr. Jarad Niemi. His support and guidance have been invaluable. Research can be daunting especially when it doesn't seem to be progressing, but Dr. Niemi's encouragement to push through and see the project out was a major factor in this finished work.

I would like to thank my committee: Dr. Emily Berg, Dr. Alicia Carriquiry, Dr. Dan Nordman, and Dr. Chong Wang. They have provided valuable mentoring throughout my time at Iowa State University and gave helpful feedback on my research to make it stronger.

Lastly, I would like to thank Jesus Christ, my wife, Breanne, and my family. Their support was crucial in my completion of the program. With their advice and affirmation, I was able to continue and finish this work. I would like to give a special shoutout to my friends I have made at Iowa State: Vianey, Anand, Danny, Nick, Nate, Manju, Katherine, and Justin. Without them my time at Iowa State would not have been nearly as enjoyable.

ABSTRACT

Influenza is a common illness which affects many people every year. In the past few years, we have seen the great impact influenza can have on the population and the health care system. For most, influenza will result in a minor inconvenience, but influenza can lead to serious health problems including death especially among the young, the elderly and expecting women. The Centers for Disease Control and Prevention (CDC) has created the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet), a network of outpatient healthcare providers throughout the United States of America and its territories who have agreed to report the weekly number of patients they see in their office showing influenza-like illness (ILI) and the total number of patients seen for any reason. Though ILINet is viewed as the gold standard for estimating influenza activity, it is often reported at a one or two week lag. Internet searches can provide a better real time view of influenza activity though they can be biased.

In this thesis, we develop multiple models using only ILINet data then develop a method for these models to incorporate a second data source through data fusion. The first model employs a Bayesian hierarchical structure with the mean modeled by an asymmetrical Gaussian functional form. Multiple hierarchical structures are compared to see which fits the data best. When forecasting, all hierarchical structures preform better than the independent model. The second model takes a functional data approach and uses functional principal component analysis to model and forecast the influenza season. Shrinkage distributions are used to choose the number of principal components. A hierarchical structure is created for the shrinkage distributions. Again, we find the hierarchical structures help in providing better forecasts. The forecasts are able to predict the peak week and peak percentage with little data from the forecasted season. Lastly, we preform a simulation study to see how

adding a second data source such as Google search data can benefit the forecasting abilities in both of these models. We found that both models can benefit from a second source of data even if it biased. The benefit is most noticeable when forecasting around the peak of the influenza season.

CHAPTER 1. OVERVIEW

Influenza is a disease which affects many people yearly around the world. It can cause a big disruption in the lives of those it inflects, but more than a disruption, influenza poses great health risks to the young, the elderly, and expecting mothers. Each year around 3-5 million cases result in severe illness and 300,000-650,000 result in death in the United States. One of the easiest way to combat influenza is through the influenza vaccine. The administration of the influenza vaccine is time sensitive. If the vaccine is administered too early, the peak effectiveness will have passed before the peak of the influenza season arrives, or if it is administered too late, it will not reach peak effectiveness before the peak of the season occurs. This makes the timing of the distribution of vaccines crucial to mitigating the effects of influenza on the population. Health officials need accurate forecasts to properly plan the distribution of influenza vaccines.

In this chapter, the contents of the other chapters in the thesis are described as well as their relation to one another. The three principal chapters are related through their use Bayesian hierarchical models to conduct inference on and forecast the influenza season.

Chapter 2 of the thesis provides a unique and flexible method to model and forecast the influenza season. A functional form of the asymmetrical Gaussian distribution is used to model the mean structure of the influenza season. The asymmetrical Gaussian functional form provides flexibility and interpretable parameters. Each parameter in the asymmetrical Gaussian functional form has an interpretation that directly relates to a facet of the influenza season. The flexibility in the functional form allows the ramp up and cool down of the season to be different as well as the starting point and end point of the season. This is crucial since the portion of the year we are interested in often does not give time to reset i.e. the starting and ending ILI percentage are not the same. Hierarchical structures based on regions and

seasons were created and compared to see whether the data presented more of a regional effect or seasonal effect. They were also compared to an independent structure with no hierarchical structure.

Chapter 3 introduces new methodology with an emphasis on forecasting influenza. This chapter uses functional principal component analysis to model and forecast the influenza season. Instead of thinking of the observations as individual weekly percentages for each region-season, the entire season for a particular region is one observation. Functional principal component analysis treats each observation as the linear combination of principal components that are derived from the data. This method provides a similar flexibility as the asymmetrical Gaussian functional form but provides better forecasts. One issue with functional principal components analysis is deciding on the number of components to use. Using a fully Bayesian approach to the problem, shrinkage distributions were assigned to the parameters for each principal component; they let the data decide which components to use and which to shrink towards 0. Multiple shrinkage distributions were evaluated; the hierarchical shrinkage distribution preformed well and showed no issues with convergence. A hierarchical structure was created for parameters in the shrinkage distributions to reflect either a regional shrinkage or seasonal shrinkage.

Chapter 4 builds on Chapters 2 and 3 and explores the possibility of using other data sources to aid in forecasting. ILINet data is viewed as the golden standard of influenza data but is often reported on a two week lag and on top of the lag there is missing data since not all doctors report their data on time. This contributes to the difference between what ILINet reports and what the influenza activity looks like right now. Internet search data can be a good secondary source of data for influenza and give a better representation of what influenza activity is right now. The problem is this other data source can be biased relative to ILINet. Using the models from Chapters 2 and 3 as the mean function, a simulation study was conducted to see how that lag affects forecasts.

Chapter 5 reviews the conclusions from the previous chapters and lists future work for improving our Bayesian hierarchical models.

CHAPTER 2. BAYESIAN HIERARCHICAL FUNCTIONAL FORM ANALYSIS OF THE INFLUENZA SEASON

2.1 Abstract

Influenza is an illness which affects many people every year. The past few seasons have highlighted the importance of being able to understand and predict the influenza season in a tmiely fashion. Influenza can lead to serious health problems including death especially among the young, the elderly and pregnant women. The Centers for Disease Control and Prevention (CDC) has created the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) to help them study influenza. In this paper, we use a Bayesian hierarchical model with an asymmetrical Gaussian functional form to model ILINet data. We compare multiple hierarchical structures as well as a vague prior model. The asymmetrical Gaussian functional form provides the necessary flexibility to capture the uniqueness in every season. A season hierarchical structure best fit the past data though a region hierarchical structure preformed the best when forecasting.

2.2 Introduction

Influenza is a common illness which many people will deal with at some point in their lives; it is an acute viral infection which attacks the respiratory system and spreads easily between people (CDC, 2017). At best, it is a minor inconvenience, but at worst, it can lead to serious health problems including death especially among the young, the elderly and pregnant women. Each year around 3-5 million cases result in severe illness and 300,000-650,000 result in death in the United States (WHO, 2018). In addition, the number of cases in the United States causes a significant economic and resource burden (M McDonnell et al., 2011; Thompson et al., 2004) on an already strained healthcare system (Inst of Medicine,

2006). Vaccines are a simple and effective way of preventing the spread of influenza. The Centers for Disease Control and Prevention (CDC) is responsible for the distribution of vaccines so they have placed significant effort on understanding influenza via increased monitoring and research. One specific goal of the research is trying to model and predict the influenza season (CDC, 2017).

Due to the importance of understanding influenza; there have been many different approaches to this problem. There are good reviews of forecasting attempts by Nsoesie et al. (2014) and Chretien et al. (2014). A significant effort has been placed on finding and using alternative data sources. The U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) created by the CDC is considered the golden standard and because of this, many researchers use this as the primary source of data. An issue with ILINet is the possible delay in reporting time of up to two weeks. This reporting delay has caused researchers to look for alternative data sources that gives a better sense of what the influenza rates are in real time. Possible alternative data sources are Google Flu Trends (GFT) (Dugas et al., 2013; Michaud, 2016; Xu et al., 2017; Cook et al., 2011; Corley et al., 2010), Twitter (Paul et al., 2014; Culotta, 2010), drug sales (Patwardhan and Bilkovski, 2012) and Wikipedia (Hickmann et al., 2015). The benefit of using these alternative data sources is that they are an open-source, readily available method of getting an estimate for the current influenza rate. Influenza rates from GFT come by aggregating search queries related to influenza or its symptoms and dividing by the total number of searches. Twitter data works in a similar way except they deal with tweets instead of searches. These alternative data sources can be a biased. Consider the time around the peak of the influenza season; there will likely be an overestimate of the influenza rate because people will be more worried about catching the flu and therefore look up its symptoms as a form of preparation. The number of searches will increase making the estimate of ILI rise, but the actual ILI percentage will not have increased as much. Another issue is that since these data sources are not funded by the government, they are pet projects for companies and can be dropped at any time. This was

the case with GFT. Google has posted that they will leave past GFT data available but that they will no longer be moving forward with the project (Google Flu Trends, 2015).

In addition to multiple data sources, a variety of models have been used: ARIMA (Soebiyanto et al., 2010; Quenel and Dab, 1998; Stroup et al., 1988), SIR (Longini Jr et al., 1986; Hall et al., 2007; Michaud, 2016) and SIR variants (Shaman and Karspeck, 2012). Historically, these models have been used to study diseases. Hierarchical structures for these models have taken advantage of the temporal structure of the data (Michaud, 2016; Yu et al., 2013a) and spatial structure (Mugglin et al., 2002). We propose a hierarchical structure based on time and location that are simple yet still beneficial to forecasts. The hierarchical structure will be implemented along with a functional form mean which will inform the general shape of the season.

2.3 ILINet

Many of the aforementioned efforts used data from the ILINet provided by the CDC due to the fact that ILINet is viewed as the gold standard for influenza surveillance (Greenspan, 2015). ILINet is a network of 2800 outpatient healthcare providers throughout the USA and its territories. These providers report over thirty-nine million yearly patient visits. The network collects data from doctors who submit the weekly number of patients they see in their office showing influenza-like illness (ILI) and the total number of patients seen that week regardless of the reason. The CDC defines ILI as "fever (temperature of $100°F[37.8°C]$ or greater) and a cough and/or a sore throat without a known cause other than influenza" (CDC, 2017). This weekly data is then aggregated into regions by taking the sum of all patients with ILI within a region and the sum of all patients seen within a region. Figure 2.1 shows how the USA is split up into ten different regions (CDC, 2011).

The USA and its territories have been split up into 10 regions decided by the CDC. Generally, the regions contain a natural grouping of states with the exception of regions 2, 9, and 10 that include states and territories that are not connected to the rest of the region.

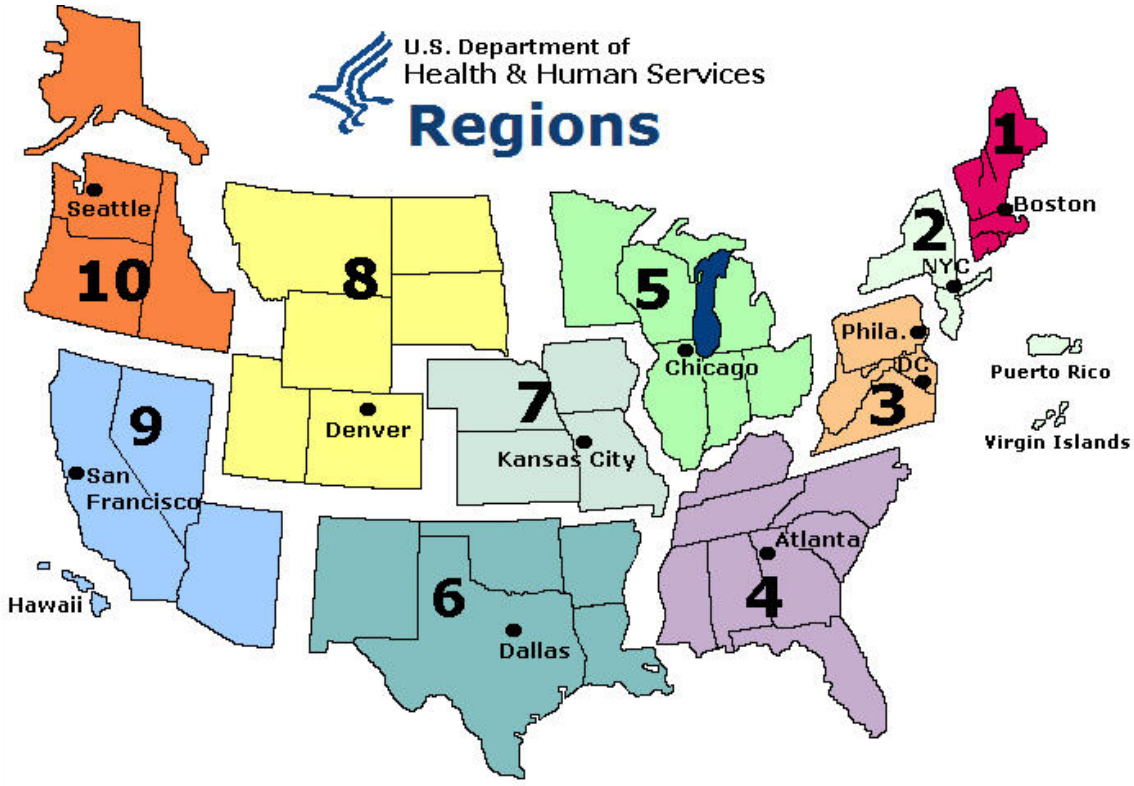

Figure 2.1 The ten regions of the United States as decided by the CDC (2011).

For this project, the seasons spanning $2006 - 2014$ were used, exclusive of $2008 - 2010$ due to H1N1. The influenza season as defined by the CDC spans Morbidity and Mortality Weekly Report (MMWR) weeks 40-20; the focus is on these weeks since these weeks see the most influenza rate activity. These weeks typically fall from November to early May. Note that in the plots of this paper the weeks range from $1 - 32/32$ such that week 1 corresponds to week 40 and so on.

The raw number of patients with ILI symptoms and the number of total patients seen were used in this study. In Table 2.1, the number of patients seen varies quite a bit, and the number of patients exhibiting ILI like symptoms is low relative to the total number of patients seen. This leads to a small range of percentages from ≈ 0 to 0.09%.

Table 2.1 Summary statistics for the total number of patients seen with ILI symptoms, the total number of patients seen, and the percentage of patients with ILI symptoms across all regions and seasons.

	Variable	Minimum	Std. Deviation	Std.Dev	Maximum
1	Num of ILI	25.00	1505.90	1435.77	11784.00
2	Total Patients	2819.00	68691.78	40057.88	165649.00
3	ILI Percentage	0.00	0.02	0.01	0.09

In Figure 2.2, the weekly percentage of patients with ILI are plotted against the weeks of the influenza season. Each region is represented as a column and each season as a row. Recall week 1 is in early November. The peaks correspond to the times normally associated with high influenza: the cold, wet, winter season. Notice that each region has its own slightly different seasonal form i.e. peak height, peak week, ramp up and cool down look different from region to region, but there is also a great seasonal effect happening on all regions. There is a great diversity in these features across each region-season.

The CDC started the Predict the Influenza Season Challenge in November of 2013 to gather and evaluate innovation and methods required to accurately forecast the influenza season (Centers for Disease Control and Prevention, 2019). The competition requires forecasters to predict the timing, peak, and intensity of the upcoming influenza season using ILINet data and any other public data available to the participants. The competition specifically asks to predict peak week; peak percentage; 1,2,3,4 week ahead forecasts; and onset. The week ahead predictions are considered short-term forecasts whereas the peak week and peak percentage forecasts are long-term forecasts. In this paper we focus on the these two types of forecasts from which you can derive the week ahead, peak week, peak percentage and onset forecasts.

2.4 Methodology

In this section, we will describe and motivate the model used to understand and forecast the influenza season. We will discuss the data model including the functional form used.

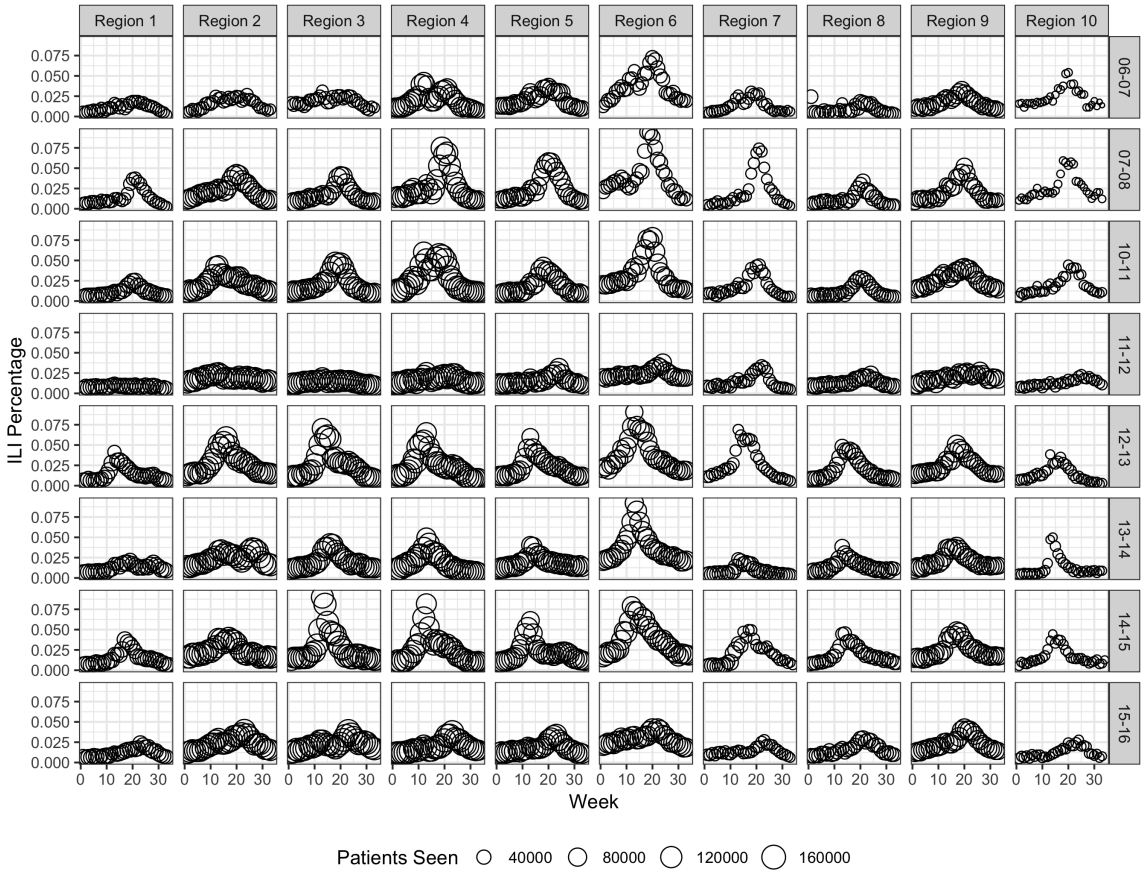

Figure 2.2 Percentage of patients that exhibited ILI faceted by seasons in the rows and faceted by region in the columns.

We will also motivate several possible hierarchical structures for the model. Lastly, the methods used to analyze the fits of the models will be presented.

2.4.1 Data Model

Let ILI_{wrs} the number of patients with ILI in week w within the season s and region r. Similarly, the total number of patients is denoted as n_{wrs}, and the true proportion of patients with influenza as ϕ_{wrs}. The data model is presented in equation 2.1.

$$ILI_{wrs} \overset{ind}{\sim} Bin(n_{wrs}, \phi_{wrs}) \tag{2.1}$$
$$logit(\phi_{wrs}) = f(w; \theta_{rs})$$

The domain of the success probability in the binomial distribution, ϕ_{wrs}, is $[0, 1]$. In order to be more flexible with our modeling choices for ϕ_{wrs}, we place it on the logit scale thereby extending the domain to the real line. The function $f()$ can take on many forms. Looking at Figure 2.2, a natural consideration would be the functional form of the normal density or the double logistic density. For this chapter, we focused on the asymmetrical Gaussian (ASG) functional form i.e. $f() = ASG()$ (Equation 2.3). The asymmetrical Gaussian is known for its flexibility and similarity to the Gaussian and double logistic.

2.4.2 Asymmetrical Gaussian Functional Form

The asymmetrical Gaussian (ASG) distribution was first introduced by Fechner in *Kollektivmasslehre* in 1897 (Wallis, 2014). The idea behind the ASG distribution, equation 2.2, was to compose a distribution from the left half of a $N(\mu, \sigma_1^2)$ and the right half of a $N(\mu, \sigma_2^2)$. Then to make sure it is a proper density it is scaled by $\frac{1}{\sqrt{2\pi}(\sigma_1+\sigma_2)/2}$ (though as seen below, it is of no use to us).

$$f_X(x) = \begin{cases} \frac{1}{\sqrt{2\pi}(\sigma_1+\sigma_2)/2} \exp[-(x-\mu)^2/2\sigma_1^2] & x \leq \mu \\ \frac{1}{\sqrt{2\pi}(\sigma_1+\sigma_2)/2} \exp[-(x-\mu)^2/2\sigma_2^2] & x \geq \mu \end{cases} \tag{2.2}$$

This gives each half of the distribution its own scaling parameter (σ_1 or σ_2). This aspect is of particular interest because it will allow the ramp up and cool down of the influenza season to be unique. For our purposes, we will not be dealing with the ASG distribution directly but will instead deal with a functional form of the ASG distribution. Since we do not need a proper distribution, we changed the form of the scaling factor allowing for

more flexibility. In the ASG functional form , equation 2.3, we swap the scaling factor for something resembling the scaling factor in model 1 from Werker and Jaggard (1997). In their paper, they define a scaling factor which allows for an initial, maximum, and ending growth rate.

$$ASG(w;\theta) = \begin{cases} \beta_1 + (\eta - \beta_1)\, exp[-(w-\mu)^2/2\sigma_1^2] & w < \mu \\ \beta_2 + (\eta - \beta_2)\, exp[-(w-\mu)^2/2\sigma_2^2] & w \geq \mu \end{cases} \tag{2.3}$$

$$\theta = (\beta_1, \beta_2, \eta, \mu, \sigma_1^2, \sigma_2^2)$$

Figure 2.3 shows the ASG functional form and what aspects of the curve each parameter corresponds to. In the curve, β_1 and β_2 stand for the baseline to be seen before and after the peak is reached; σ_1^2 and σ_2^2 control the ramp up and cool down of the curve, respectively; μ represents the peak week i.e. the week in which the curve reaches its peak; and η represents the peak height.

2.4.3 Hierarchical Structures

Now that the data model and functional form of the mean have been established, the model for the θ_{rs} parameters needs to be established. One possible idea is to model the parameters within θ_{rs} individually and independently, but if they are not independent, we would be throwing away information. Taking this into consideration, a possible appropriate modeling choice for θ_{rs} would have it follow a multi-variate normal distribution. This will allow for the data to inform us of the correlation in the parameters through the variance-covariance matrix.

$$\theta_{rs} \overset{ind}{\sim} N(\theta, \Delta\Omega\Delta)$$

In order to do this, the σ_1 and σ_2 parameters in the ASG functional form are logged so that all the parameters lie on the real line. We will model the variance-covariance matrix

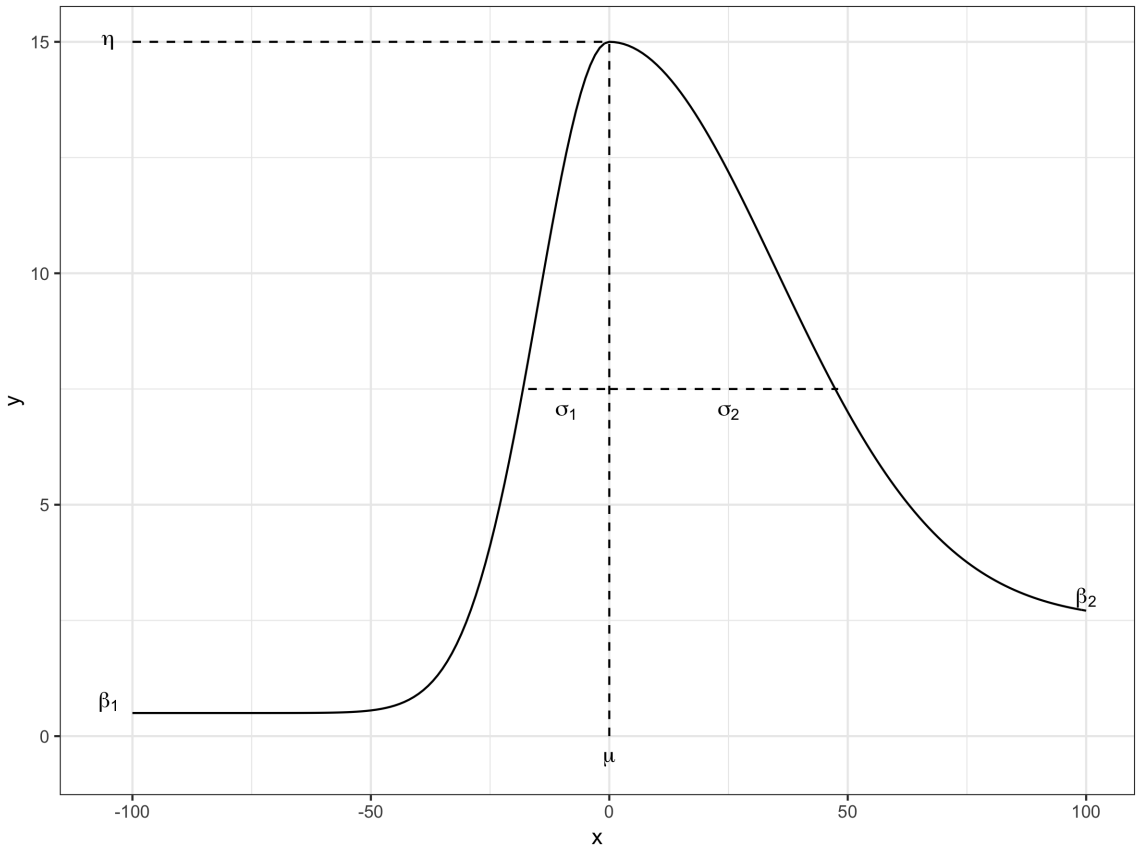

Figure 2.3 The asymmetrical Gaussian functional form and its parameters. The parameters placed according to the corresponding aspects they control.

using the separation strategy proposed by Barnard (Barnard et al., 2000). This approach decomposes the variance-covariance into a diagonal matrix of standard deviations (Δ) and a correlation matrix (Ω) i.e. $\Sigma = \Delta\Omega\Delta$. The decomposition is recommended in the `Stan` manual (Stan Development Team, 2018) and provides independence a priori on the standard deviations and the correlation matrix.

Now, let us address the question whether there should be any structure on the θ, Δ, and Ω parameters. In Figure 2.2 there were clear differences in each regions. The peak heights were different across regions, even when the peak weeks were consistent. One consideration

is to set up the means and standard deviations so that each θ_{rs} in a region can borrow information across seasons.

$$\begin{aligned}
\theta_{rs} &\overset{ind}{\sim} N(\theta_r, \Delta_r \Omega \Delta_r) \\
\theta_r &\overset{ind}{\sim} N(\theta, \Delta \Omega \Delta) \\
\Delta_r &= diag(\varsigma_{r,1}, \cdots, \varsigma_{r,6}) \\
\Delta &= diag(\varsigma_1, \cdots, \varsigma_6)
\end{aligned} \tag{2.4}$$

This structure will be known as RegMnSd (equation 2.4) since the mean (θ_r) and the standard deviations (Δ_r) are region dependent. Reg stands for region, Mn stands for mean and Sd represents standard deviation. For example in region 1, we are borrowing information across the eight seasons ($\theta_{1,1}, \theta_{1,2} \cdots \theta_{1,8}$) to learn the region 1 mean, θ_1.

In Figure 2.2, there is also a visible seasonal pattern. The data points in the regions within a single season are very similar; potentially more so than they are with past seasons in the same region. Taking this we can create a hierarchical structure where we borrow information across regions to learn the season mean (θ_s) and standard deviation (Δ_s). This structure will be known as SeasMnSd (equation 2.5) where Seas stands for season and Mn and Sd maintain the same meanings from before.

$$\begin{aligned}
\theta_{rs} &\overset{ind}{\sim} N(\theta_s, \Delta_s \Omega \Delta_s) \\
\theta_s &\overset{ind}{\sim} N(\theta, \Delta \Omega \Delta) \\
\Delta_s &= diag(\varsigma_{s,1}, \cdots, \varsigma_{s,6}) \\
\Delta &= diag(\varsigma_1, \cdots, \varsigma_6)
\end{aligned} \tag{2.5}$$

Looking at the previous hierarchical structures it is easier to conceive of borrowing information for the mean than borrowing information for the standard deviations. An easy simplification can be made by removing the region or season dependence on the standard deviations. This structure suggests that the way the different θ_{rs} vary from either their

season or region means is the same across all seasons and regions. These will be known as SeasMean (equation 2.6) and RegMean (equation 2.7) using the same naming conventions above.

$$\begin{aligned} \theta_{rs} &\overset{ind}{\sim} N(\theta_s, \Delta\Omega\Delta) \\ \theta_s &\overset{ind}{\sim} N(\theta, \Delta\Omega\Delta) \\ \Delta &= diag(\varsigma_1, \cdots, \varsigma_6) \end{aligned} \tag{2.6}$$

and

$$\begin{aligned} \theta_{rs} &\overset{ind}{\sim} N(\theta_r, \Delta\Omega\Delta) \\ \theta_r &\overset{ind}{\sim} N(\theta, \Delta\Omega\Delta) \\ \Delta &= diag(\varsigma_1, \cdots, \varsigma_6) \end{aligned} \tag{2.7}$$

Another variation on the hierarchy structure would be to have the mean be region dependent and the standard deviation be season dependent (RegMnSeasSd, equation 2.8). This structure argues that the functional form for a given region will more or less remain the same but it can shifted around depending on whether the spread of influenza in a particular season is mild or rampant.

$$\theta_{rs} \overset{ind}{\sim} N(\mu_r, \Delta_s\Omega\Delta_s) \tag{2.8}$$

Likewise, we could consider the opposite happening in which the mean is season dependent and the standard deviation region dependent (SeasMnRegSd, equation 2.9).

$$\theta_{rs} \overset{ind}{\sim} N(\mu_s, \Delta_r\Omega\Delta_r) \tag{2.9}$$

Lastly, we considered an independent model (equation 2.10). This model does not borrow information and naively learns θ_{rs} from the current season.

$$\theta_{rs} \overset{ind}{\sim} N(m_0, C_0) \tag{2.10}$$

2.4.4 Forecasting

The focus of the CDC's competition is forecasting upcoming influenza seasons to a consistent high degree of accuracy in hopes that public health policy can be based on these forecasts. In order to see how the proposed models forecasting performed, short and long term forecasts were analyzed. Week ahead forecasts are a form of short-term forecasts that allow us to see what will be happening in the near future. To create week ahead forecasts using m weeks in the forecasted season, the model is run as normal including m weeks in the season to be forecasted. Then the parameter posterior samples are plugged into the asymmetrical Gaussian functional form (Equation 2.3) and the week ahead forecasts emerge from looking at weeks following week m. A 1 week ahead forecast from week 10 would include 10 data points from the forecasted season and would forecast what will happen in week 11. We can create a whole season forecast of 1 week ahead forecasts by creating 1 week ahead forecasts by including 1 week in the forecasted season, then including 2 weeks in the forecasted season, and so on until a season forecast is created of 1 week forecasts. Two, three, and four week ahead forecasts are created in a similar fashion.

While short term forecasts are definitely beneficial, there is also a need for long-term forecast. Long-term forecast give insight to what the overall season is going to look like. Of particular interest is the peak timing and height of the forecasted influenza season. To create long-term forecasts, we included 3, 10, 15, or 20 weeks of the new season and all of the past seasons to forecast the rest of the new season like the week ahead forecasts. The models were ran with these different data sets, the parameter posterior samples were plugged into the asymmetrical Gaussian functional form, and the forecasts are created for the entire season.

2.4.5 Model Checking

In this subsection, the methods used to check how the model fits will be listed. First we will list the necessary detail to fit the models. Next, we discuss how we assess convergence

of the Markov chain Monte Carlo (MCMC). Then we describe the methods used to check if the asymmetrical Gaussian functional form's flexibility is necessary. Lastly, the methods used to check the model fit and forecasting fit are listed. These methods will help us discern which hierarchy structure will fit the data best.

2.4.5.1 Estimation

The priors that need to be specified in this model are minimal; they are listed in equation 2.11.

$$\begin{aligned} \varsigma_{r,i}, \varsigma_{s,i} &\overset{ind}{\sim} t_4^+(a,b) \\ \varsigma_i &\overset{ind}{\sim} t_4^+(c,d) \\ \theta &\overset{ind}{\sim} N(m_0, C_0) \\ \Omega &\overset{ind}{\sim} LKJ(\nu) \end{aligned} \tag{2.11}$$

Priors are needed on the mean, θ, the standard deviations, ς, and the correlation matrix, Ω. The prior for θ is a combination of independent normal priors for the parameters in θ. The intent was to choose mean and covariance matrix priors, m_0 and C_0, that are non-informative. Priors of $m_0 = (0, 0, 2, 15, 2, 2)$ and C_0, a diagonal matrix of $(1, 1, 0.75, 2, 0.5, 0.5)$, provide a vague starting curve. Figure 2.4 shows the asymmetrical Gaussian functional form of m_0 in bold and a few random draws from $N(m_0, C_0)$ in lighter grey. The bold line gives a pretty standard looking curve and the random draws show the variety of curves you can get from the variance.

The standard deviations get similar standard priors in the form of independent standard half-t distributions with 4 degrees of freedom. The correlation matrix provides less obvious choices. In this analysis, the LKJ prior was used. This correlation matrix prior was proposed by Lewandowski (2009) and has one parameter, ν, which controls the shrinkage in the correlation matrix. The correlation matrix shrinks towards the identity as ν gets large. By

taking ν to be 1, the prior is uninformative in what form the correlation matrix takes. For $\nu < 1$ and shrinking, the prior favors more correlation. In our study we set ν equal to 1.

The models were fit using Hamiltonian Monte Carlo via `Stan` (Stan Development Team, 2016) through `R` (R Core Team, 2016). One chain was ran for $10,000$ iterations; 5000 of which were warm-up iterations. Since models with non-linear means can be troublesome, maximum likelihood estimates were used as starting values.

2.4.5.2 Convergence Check

Once the model is fit, to assess for issues with non-convergence, the Geweke diagnostic was used. Geweke (1992) created a convergence diagnostic for Markov chains where if there are no obvious causes for concern, the statistic should look like draws from a standard normal distribution.

Trace plots will also be used for a visual inspection of convergence issues.

2.4.5.3 Functional Form Check

While flexibility can be good, it can also add more model complexity and computational complexity. It is important to examine the need for flexibility. To do this, we compared functional form to that of a normal distribution. These can be compared easily since the normal functional form is a special case of the asymmetrical Gaussian functional form where $\beta_1 = \beta_2$ and $\sigma_1^2 = \sigma_2^2$. One can check for evidence of this by using posterior probabilities listed in equation 2.12.

$$\begin{aligned} p(\beta_1 > \beta_2|y) &= \frac{1}{n.iter} \sum_{i=1}^{n.iter} I\big(\beta_1^{(i)} > \beta_2^{(i))}\big) \\ p(\sigma_1 > \sigma_2|y) &= \frac{1}{n.iter} \sum_{i=1}^{n.iter} I\big(\sigma_1^{(i)} > \sigma_2^{(i)}\big) \end{aligned} \tag{2.12}$$

If $p(\beta_1 > \beta_2|y)$ is near 1 or 0, this is evidence that the β_1 and β_2 are different and necessary. If the posterior probability hovers around 0, that signals that the difference is not great and we do not need two intercepts. The same reasoning holds with $p(\sigma_1 > \sigma_2|y)$.

2.4.5.4 Model Fit

Many hierarchical structures have been proposed, but which hierarchical structure best represents the data? This will be checked by comparing mean squared error (MSE), widely applicable information criterion (WAIC), and mean square forecast error (MSFE). The formulas are listed in equation 2.13.

$$\begin{aligned} MSE &= \frac{1}{n}\sum_{i=1}^{n}(\frac{ILI_{wrs}}{n_{wrs}} - \widehat{\phi_{wrs}})^2 \\ MSFE &= \frac{1}{n}\sum_{i=1}^{n}(\frac{ILI^*_{wrs}}{n^*_{wrs}} - \widehat{\phi^*_{wrs}})^2 \\ WAIC &= \widehat{\text{elpd}}_{waic} \\ &= \widehat{\text{lpd}} - \hat{p}_{waic} \end{aligned} \tag{2.13}$$

MSE allows us to see how far away the within sample predictions from the model are to the actual data whereas MSFE shows how far away the out of sample predictions (forecasts) are, but neither account for model complexity. WAIC was introduced by Watanabe (2010) as an alternative to DIC for Bayesian analysis. Though DIC has become popular with its usage through BUGS and JAGS, it can be difficult to properly calculate as shown in Celeux (2006). WAIC offers an advantage to Bayesian models because it is was created with Bayesian analysis in mind. As many have pointed out, DIC has issues with Bayesian models due to it being based on a point estimate (Linde, 2005; Plummer, 2008; Vehtari and Gelman, 2014) rather than taking the full posterior into consideration. In certain contexts, WAIC can be used to compare fit amongst models (Vehtari et al., 2017).

2.5 Analysis of ILINet

2.5.1 Convergence Check

To assess for non-convergence, the Geweke diagnostic was used. Figure 2.5 plots the Geweke diagnostic statistics for each parameter in the asymmetrical Gaussian functional

form for all region-season combinations where each hierarchical structure has a different linetype. Looking at the figure, there are no causes for concern; the statistics look as though they are random draws from a normal distribution.

Another method of assessing convergence issues is to look at how the prior influenced the posteriors. To assess whether the posteriors are being controlled by the priors, Figure 2.6 plots the posterior samples of the parameters of θ and overlays their priors. The posterior densities have centered away from the center of the prior densities. This is good evidence that our priors are not dominating the posteriors.

Trace plots were also looked at and though not included here, there was no cause for concern.

2.5.2 Model Fit

After checking for convergence issues with the MCMC, we checked if the asymmetrical Gaussian functional form is necessary and useful. Figure 2.7 shows the posterior probabilities outlined in equation 2.12 for all seasons ad regions and all hierarchical structures. Most of the probabilities are along the corners which suggest that the β and σ parameters are indeed different. Consequently, the flexibility which the asymmetrical Gaussian distribution provides is necessary.

The flexibility in the asymmetrical Gaussian functional form also allows us to look at possible trends from season to season. A natural question would be if the estimates for the curves change from season to season. Figure 2.8 plots the estimates of the parameters that define the ASG curve each region-season. The parameter posterior mean estimates change with the season. The spread of the estimates generally remains the same, but shifts up and down as the seasons progress with the exception of β_1 and β_2 which do not shift much. This is evidence that the season more so than the region effects the trend of the influenza season. This could be used as an argument for a season hierarchical structure.

As expected, β_1 and β_2 did not change much from season to season though we are still interested in looking at the minor changes in baselines. Figure 2.9 shows the mean posterior estimates and the 95% credible intervals of β_1 and β_2 for all seasons for Region 1 with the actual first and last data points of the season plotted as the black points. In general, the estimates for β_2 are lower than the actual last point of the season. At the beginning of the season the β_1 parameters are basically the same implying that the baselines are not changing from season to season. This is consistent with what we know from the CDC baselines for each region. It also suggests that if the entire year was being modeled rather than the 40 weeks in which the CDC is interested, the same intercept could be used on both halves of the asymmetrical Gaussian distribution. Since we are primarily interested in 32 weeks of the year, having two intercepts gives the necessary flexibility. This plot is for Region 1, but the ideas hold for the rest of the regions.

After noting the need for the flexibility in the asymmetrical Gaussian functional form, model fit was considered and comparisons of the different hierarchical structures. Figure 2.10 shows the posterior mean fits and their 95% credible intervals for all hierarchical structures on all regions in the 2012-2013 season. With the exception of the common model, all the model fits are similar. It is surprising that the independent model performs just as well as the hierarchical structures and that the credible intervals are not noticeably bigger.

Table 2.2 The different hierarchical structures across MSE and WAIC; the minimum value of each statistic has been subtracted from each row to make comparisons easier.

	Common	Indep	RegMn	RegMnSd	SeasMn	SeasMnSD	RegMnSeasSd	SeasMnRegSd
Δ MSE	837205.79	115.67	123.74	64.80	104.68	60.53	54.19	0.00
Δ WAIC	957314.97	93.38	156.22	0.00	147.31	370.03	57.14	1057.44

Beyond a visual inspection, MSE and WAIC were used to compare the fits of the hierarchy structures. Table 2.5.2 shows the change in MSE and WAIC for all the hierarchical structures; the lowest MSE and WAIC have been subtracted from the rest of the MSE and WAIC values, respectively. In general, the region hierarchical structures had the lowest

WAIC values and the seasonal hierarchical structures had the lowest MSE values. The structure with a season mean and region standard deviation has the lowest MSE but the lowest WAIC comes from simpler region mean and region standard deviation model. The two statistics offer competing ideas for the best fit, but looking ahead to forecasting, we should favor a model with fewer parameters that provides a good fit.

2.5.3 Forecasting

One of the main points of focus for studying influenza seasons is being able to forecast an upcoming influenza season. In order to see how the proposed models forecasting performed, we created both short and long term forecasts. 1, 2, 3, and 4 week forecasts are presented in Figures 2.11, .7, .8 and .9, respectively. These plots show the week ahead forecasts for the 2016-2017 influenza season. One thing to notice in the plots is that the number of early forecasted peaks grows as the forecasts get further out and the number of wide error bounds also increases. The further out we predict, the less sure we should be about those predictions. This happens in long-term predictions as well.

While the plots give a general sense of how the forecasting is doing, it is difficult to compare the different hierarchical structures with any sort of precision. Mean square forecasting error (MSFE) will be used to compare the forecasting abilities of the different structures. This is the same formula used for MSE except the week ahead forecast is now the estimate. The MSFE for each hierarchical structure is presented in Table 2.5.3 with the lowest MSFE for each row subtracted from the rest and then multiplied by $1x10^6$ to make comparisons easier. The RegMn hierarchical structure has the lowest MSFE for all week ahead forecasts though the SeasMn structure is close. This may seem contrary to our results earlier where the seasonal pattern seemed to dominate the regional aspect but in these forecasts there is not enough data in the forecasted season to learn the forecasted season mean well.

While short term forecasts preformed fairly well so now we will examine the long term forecasts. Figure 2.12 shows the forecast of all hierarchical structures for the $2016 - 2017$

Table 2.3 Mean square forecast error (MSFE) for all hierarchical structures' week ahead prediction; the minimum value of each statistic has been subtracted from each row and has been multiplied by $1x10^6$ to make comparisons easier.

Weeks Ahead	Indep	RegMn	RegMnSd	RegMnSeasSd	SeasMn	SeasMnRegSd	SeasMnSd
Δ 1 Week Ahead MSFE	9.09	0.00	12.08	23.74	7.84	3.31	3251.51
Δ 2 Week Ahead MSFE	63.45	0.00	38.89	186.78	19.86	30.50	6272.22
Δ 3 Week Ahead MSFE	253.21	0.00	101.81	1464.50	44.88	306.45	3411.15
Δ 4 Week Ahead MSFE	812.58	0.00	232.50	4415.74	80.97	2368.39	4192.70

influenza season using only 3 weeks of the new season. The benefit of the hierarchical structures are seen in the forecasts. While the independent model seemingly performs as well as the hierarchical structures, when all the data is present, the hierarchical structures helps inform the forecast when the data is not present.

When the independent model is removed as in Figure 2.13, the nuances of the other forecast can be seen. The mean forecasts are all fairly similar though the credible intervals for the forecast can vary quite a bit. For example, in Regions 2 & 4, the credible intervals for the models RegMnSd, SeasMnSd and SeasMnRegSd are quite large.

This happens because the posterior densities for the parameters in the asymmetrical Gaussian functional form have large uncertainty. This is especially true for the β_2 parameter which controls the second intercept. Figure 2.14 shows the posterior densities of the standard deviation of the parameters in the asymmetrical Gaussian functional form for the hierarchical model using a region mean and standard deviation structure while using 3 weeks in the forecasted season. We can see that η, μ and β_2 have high standard deviations, especially for Regions 2 & 4. This contributes to the wide prediction intervals.

It makes sense that with only 3 weeks of data, the forecasts would rely on the past. With so little information about the current season, there is nothing to do but rely on the past. We are interested in what the forecast will do when it is either at the beginning or in the middle of the ramp up period of the season. The 10 week and 15 week forecasts will help answer this question. While the 3 week forecast relied quite a bit on the priors and past seasons for its forecasts, there was a curious pattern in the 10 and 15 week forecasts.

The forecast would basically drop off whenever the included weeks dropped off. The general pattern is that the forecast would consider the peak of the season one or two weeks after the data stopped then it would continue the cool down process. Figure 2.15 shows this phenomenon. There is still the issue with prediction intervals blowing up except that now it is occurring right after the data stops in most cases whereas before is was happening towards the end of the forecast.

Even with poor forecasts, the benefit of the hierarchical structure again shows up with the 10 and 15 week forecasts. The independent model still follows the prior as soon as the data stops. This highlights the benefit of the using a hierarchical structure because while it does not forecast perfectly, it is able to forecast better than a model using no information of past seasons.

To check which hierarchical structure forecasts best long-term, MSFE was looked at. Figure 2.16 shows the MSFE for the mean forecast and the MSFE treating the 95% credible intervals as forecasts for the ten week forecast. The models with high MSFE for the upper credible interval are where the interval blows up. This tends to be the case when standard deviations are allowed to change which suggests using either the RegMn or SeasMn structure. Having region means helps in a couple cases to provide better forecasts suggesting the RegMn structure. This makes sense because with minimal information about the current season, it will be difficult to determine a season mean.

2.6 Discussion

In this paper, we proposed hierarchical structures that are simple and beneficial to forecasting. It outperforms the independent model in almost all week ahead forecasting. The asymmetrical Gaussian functional form was very good at capturing the mean of the influenza season. It provided the needed flexibility to allow for the variations in a season; this was shown in Figure 2.7. The posterior probabilities show that the intercepts (β_1, β_2) and the deviations (σ_1,σ_2) are different from each other revealing the need for the flexibility.

Though this may not be the case if the entire year is modeled and not only the influenza season.

When modeling the past seasons, MSE favors the more complex models and WAIC favors simpler models. Season hierarchical structures preformed better at modeling past seasons, but it is not clear how much better. For example, the independent model performs fairly close to the hierarchical structures and takes considerably less computation time. For strictly learning about the past history, it may be better to forgo the hierarchical structure in favor of computation time. Although, the parameter posterior mean estimates of the curve show a seasonal pattern, so it may be wise to keep a season hierarchical structure even if the independent model performs as well. In forecasting, the benefit of the longer computation time for using a hierarchical structure is shown. Generally, simpler region hierarchical structures perform better.

The importance of hierarchical structures is shown in both the long-term and short-term forecasts. The hierarchical structures helped the forecasts be more realistic than the independent model. Long-term forecasts still had the issue of the peak occurring much too early. More work is needed to give accurate long term forecasts. Certainly, we care deeply about what will happen in the coming few weeks, but there is also interest in the timing of peak and how high that peak will reach. More focus on long term forecasts will help answer these important questions.

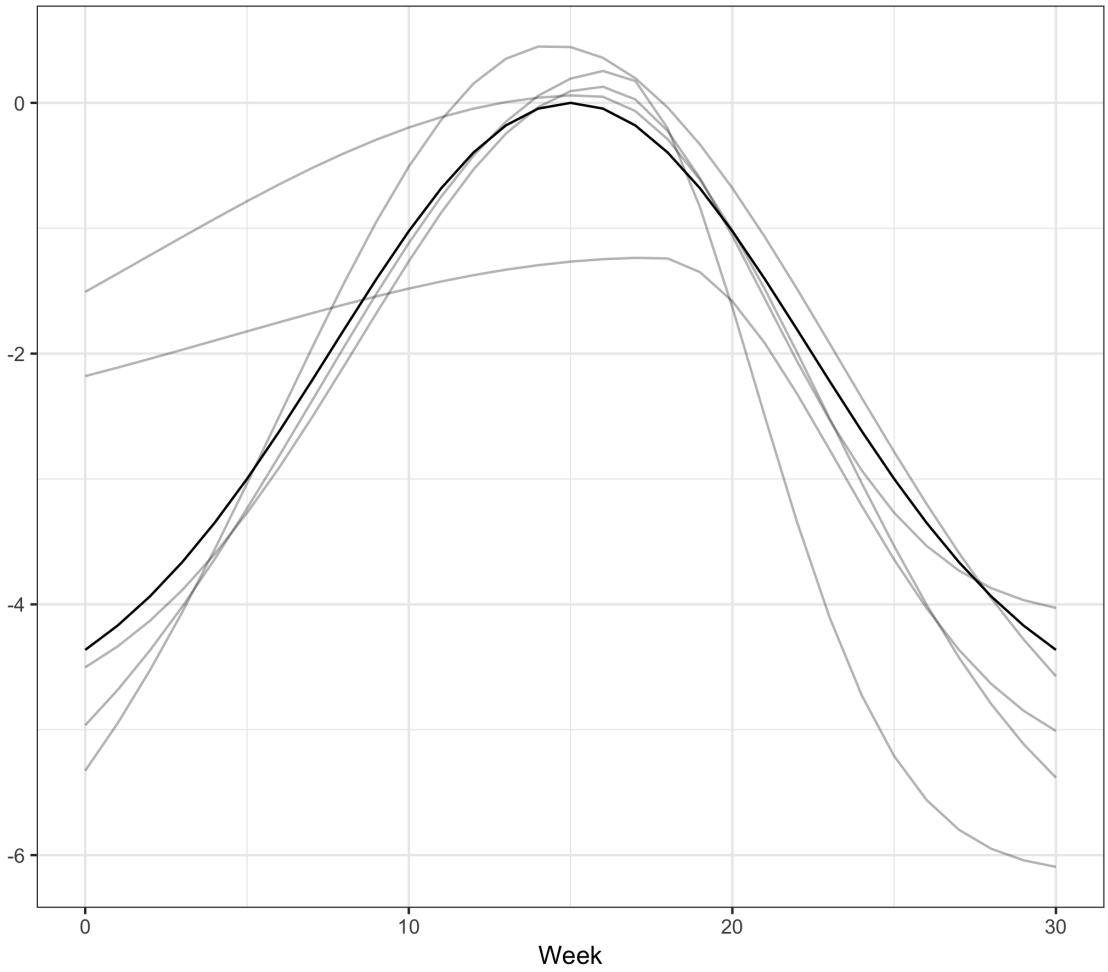

Figure 2.4 The base Asymmetrical Gaussian form from the mean and covariance of the prior. The bold line is the plot using the mean parameters and the lighter shaded lines are random draws from $N(m_0, C_0)$.

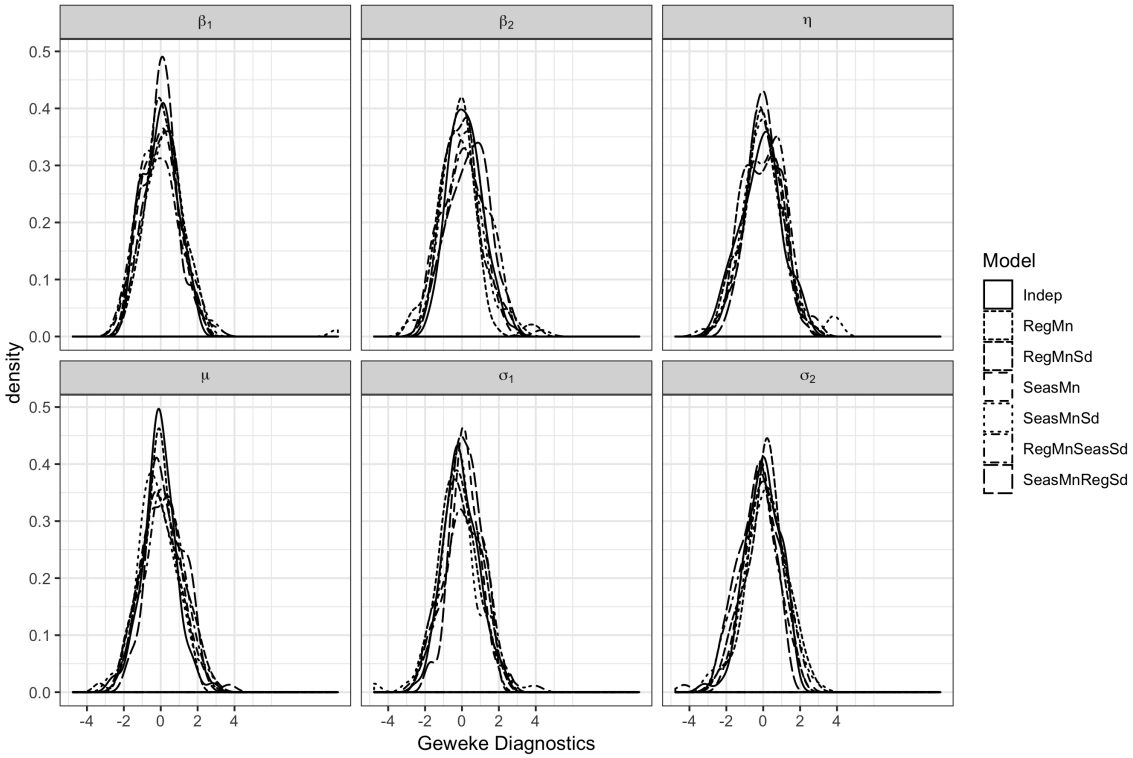

Figure 2.5 Densities of the Geweke diagnostics for each of the parameters in the asymmetrical Gaussian function over all the season-regions and the 6 different hierarchical structures.

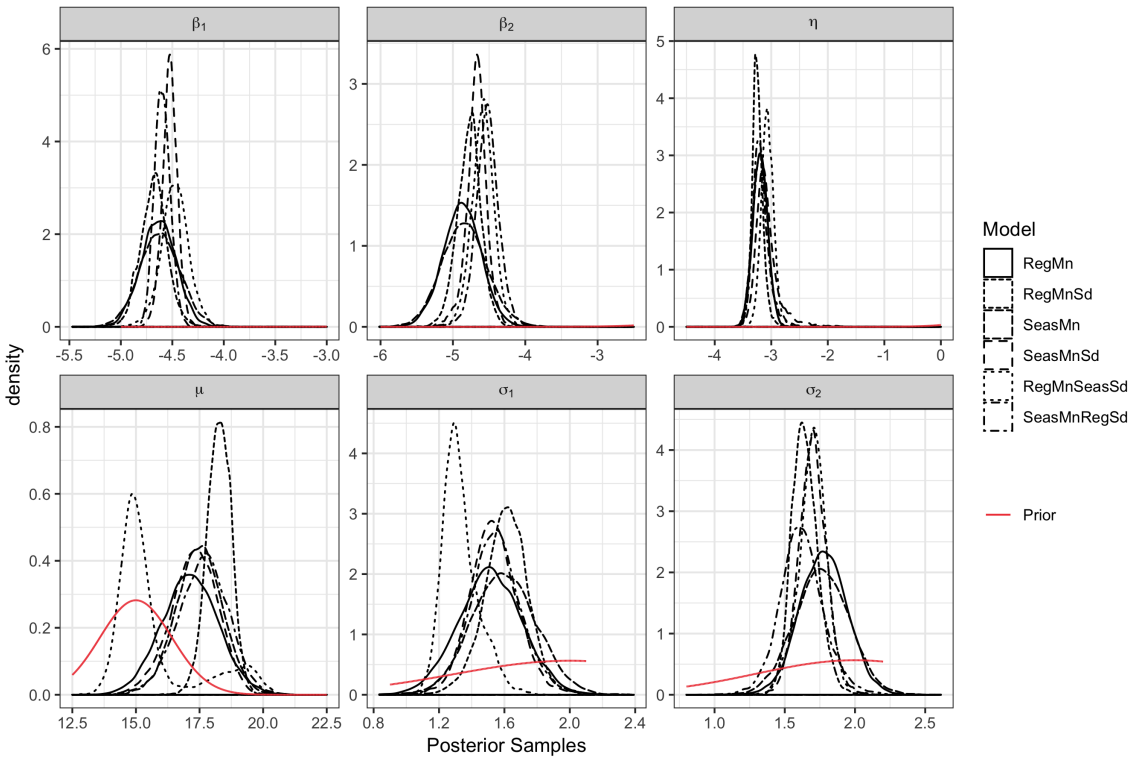

Figure 2.6 Posterior densities for the parameters of θ for the different hierarchical structures with the prior plotted in red.

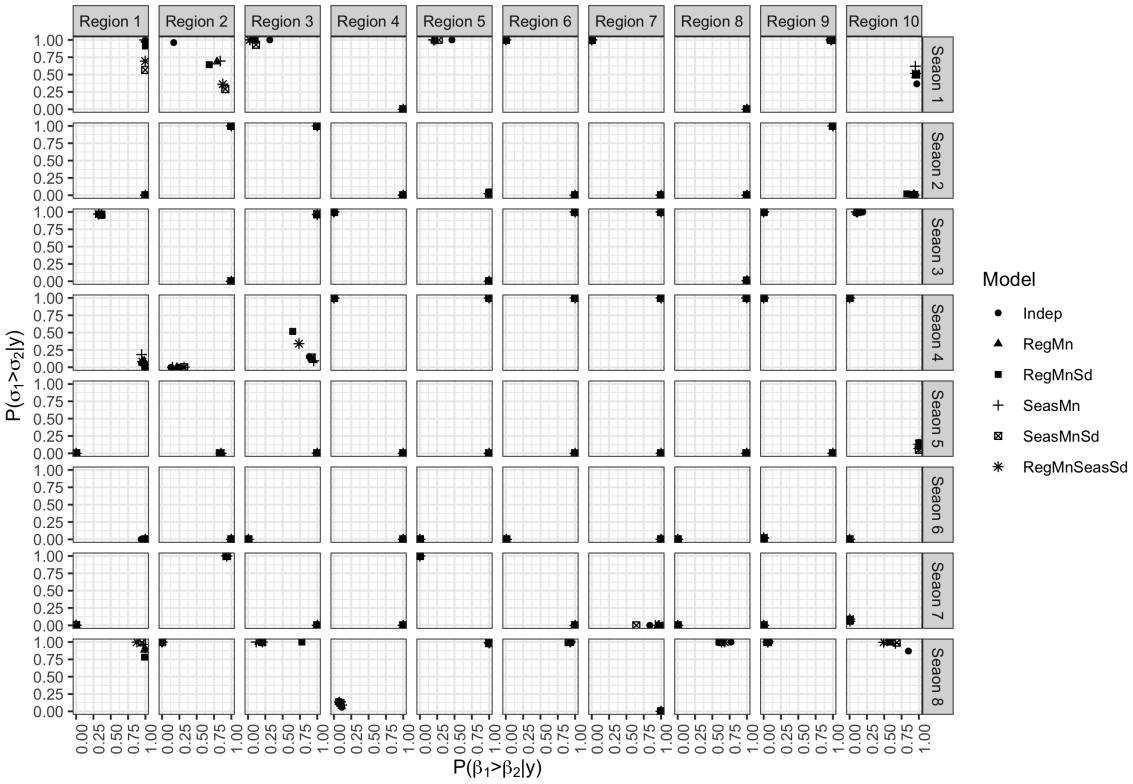

Figure 2.7 Posterior probabilities that $\beta_1 > \beta_2$ and $\sigma_1^2 > \sigma_2^2$ all season-region combinations for all models. Most of the points are in the corners which is evidence that the parameters are not equal. This shows the need for the flexibility of the asymmetrical Gaussian distribution.

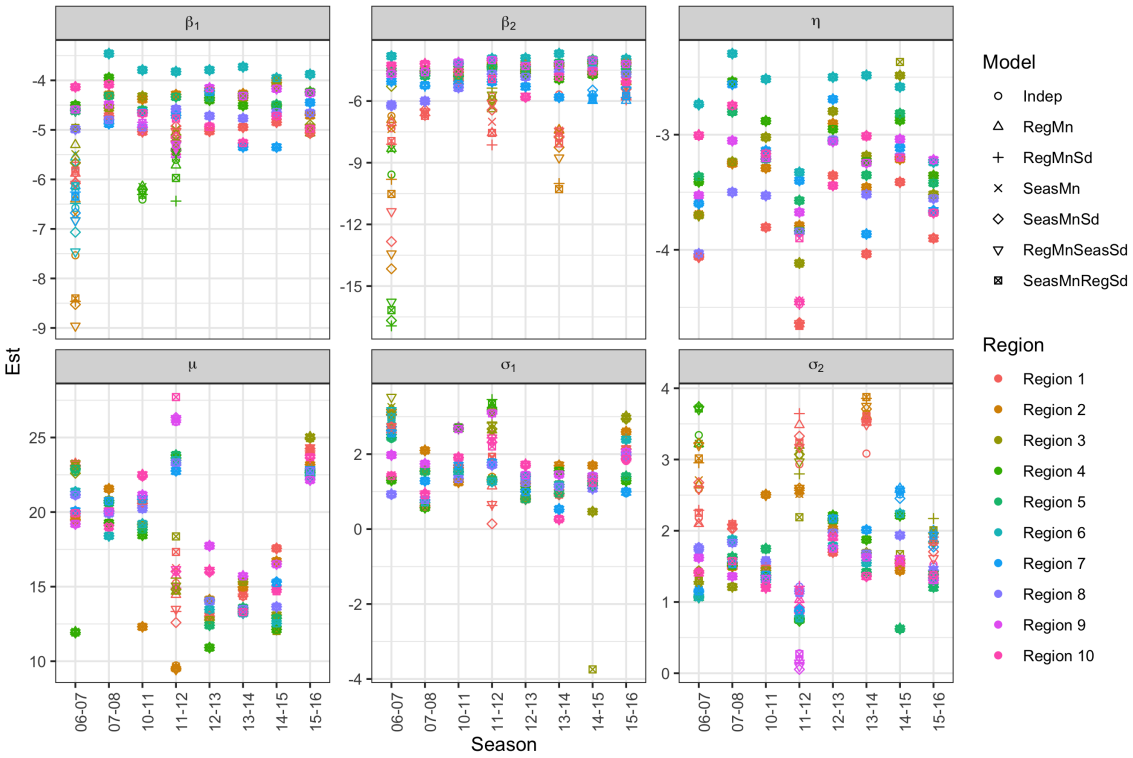

Figure 2.8 Posterior mean estimates for each parameter in the asymmetrical Gaussian functional form for all hierarchical structures in all regions and seasons. The colors represent the different regions and the shapes represent the hierarchical structures.

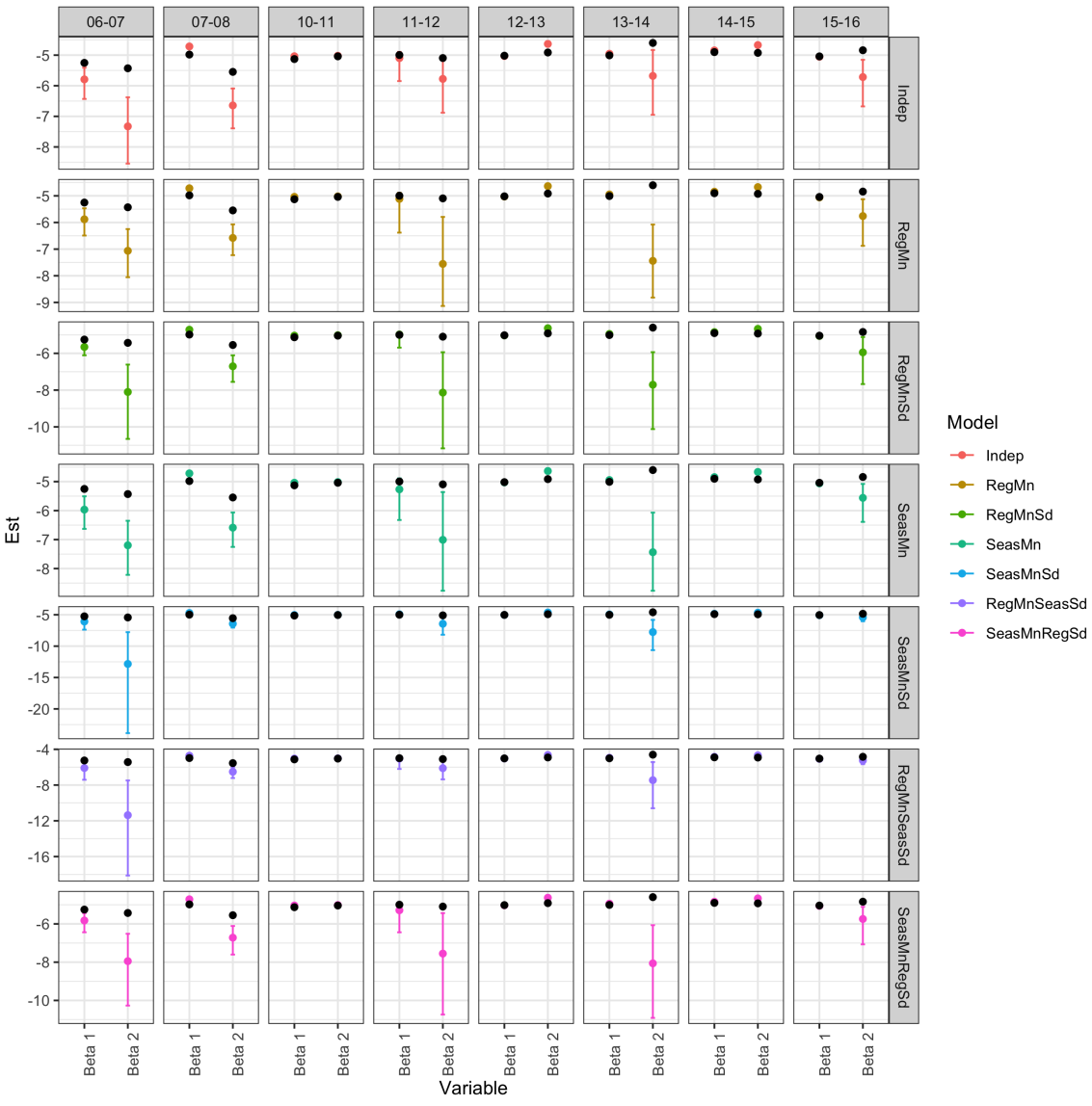

Figure 2.9 Posterior mean estimates and the 95% credible intervals of β_1 and β_2 for all seasons for Region 1 with the actual first and last data points of the season as the black points.

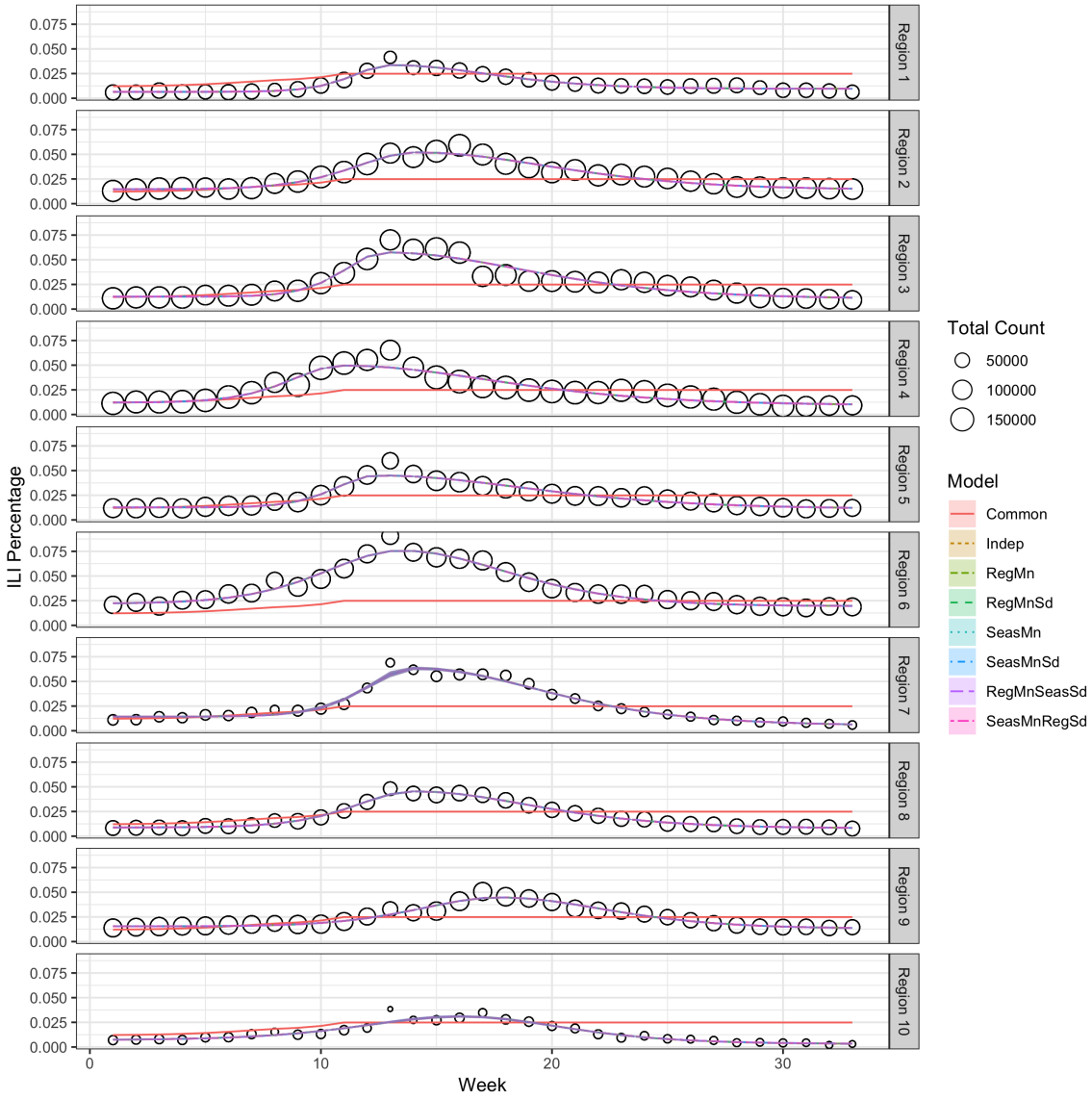

Figure 2.10 Posterior mean fit and 95% credible intervals of all hierarchical structures on the ten regions in the 2012-2013 season.

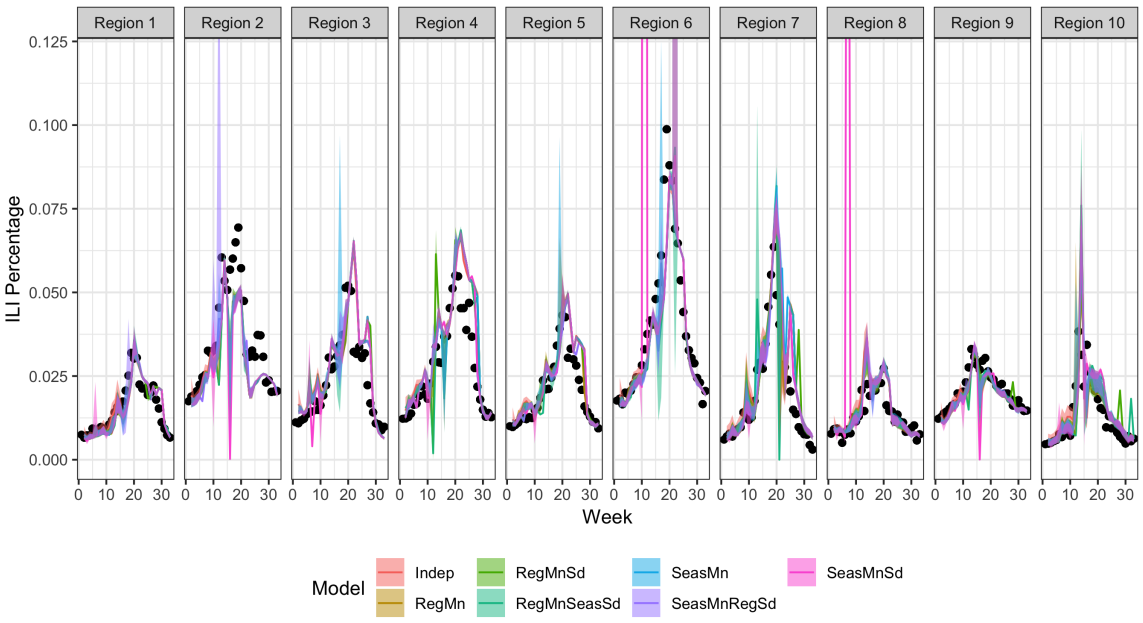

Figure 2.11 1 week ahead forecasts and their 95% credible intervals for all regions in the $2016 - 2017$ influenza season.

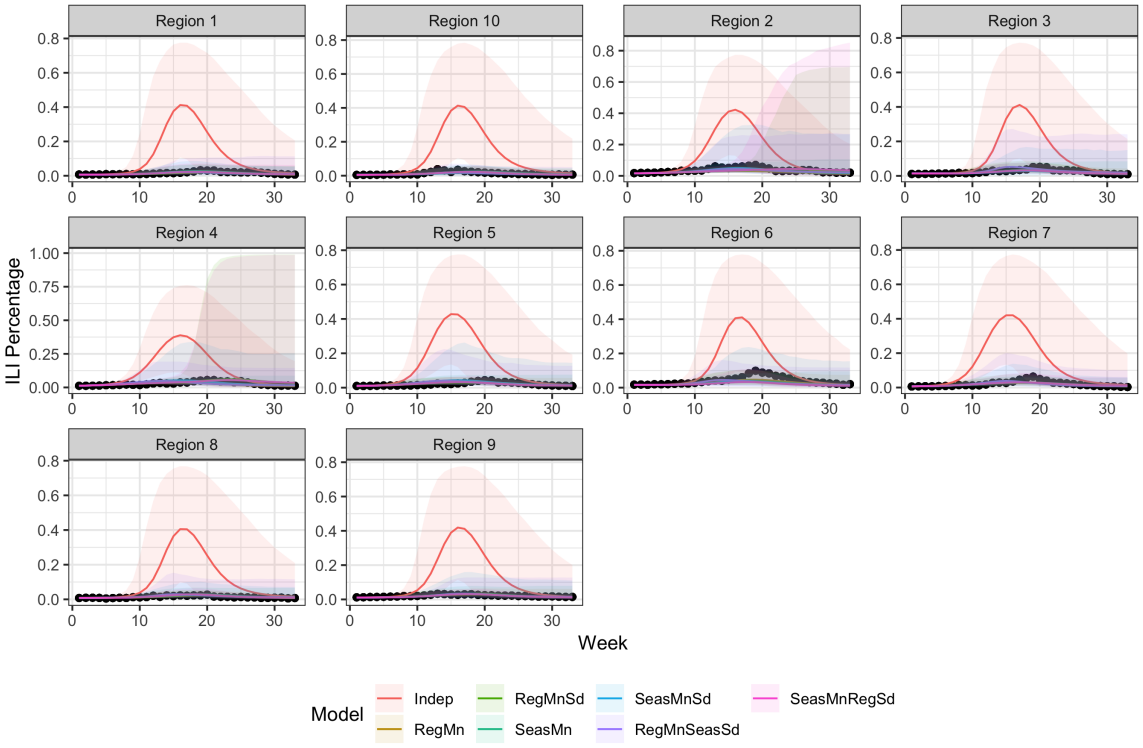

Figure 2.12 Long-term forecasts of all hierarchical structures for all regions in the 2016-2017 influenza season including only 3 weeks of data from the forecasted season.

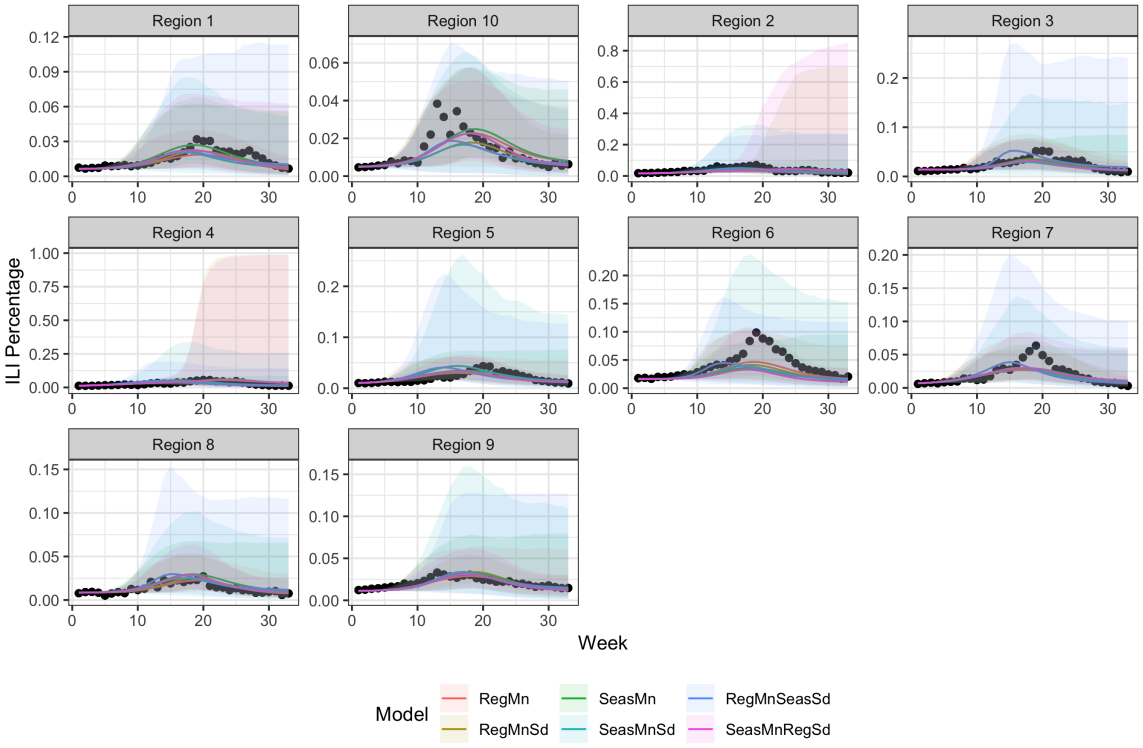

Figure 2.13 Long-term forecasts of all hierarchical structures except the independent model for all regions in the 2016-2017 influenza season including only 3 weeks of data from the forecasted season.

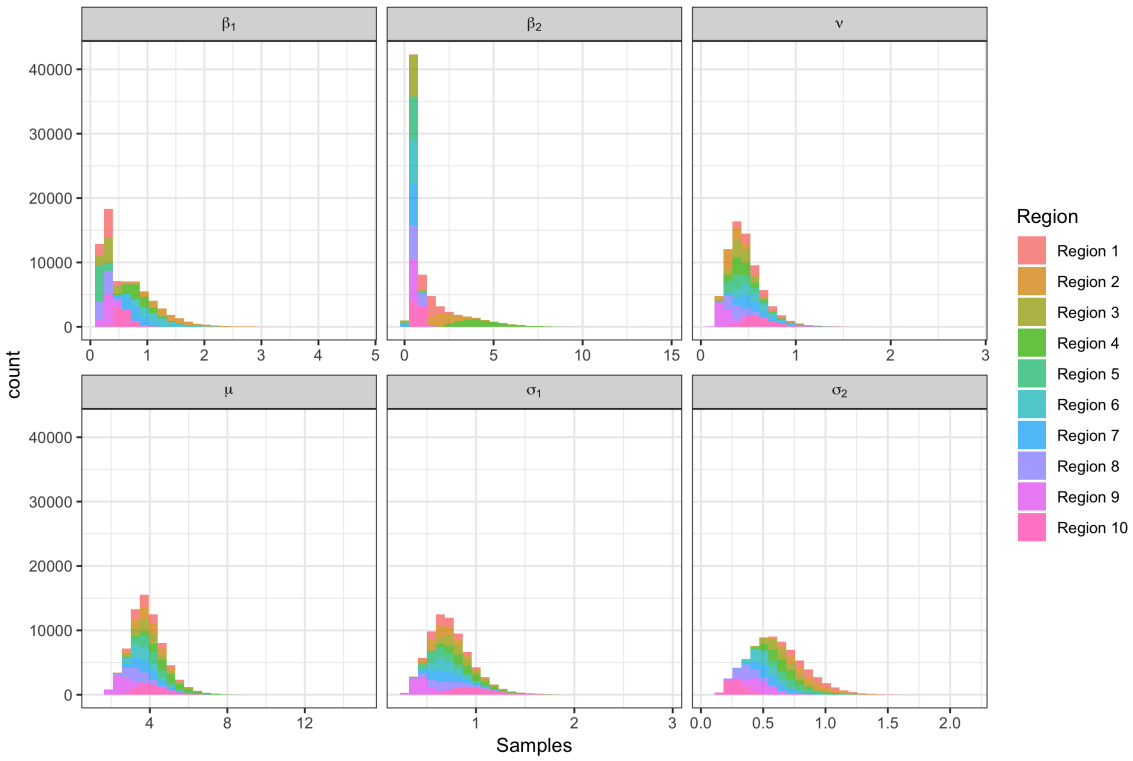

Figure 2.14 Posterior densities of the standard deviations of the parameters in the asymmetrical Gaussian functional form for the hierarchical structure using a region mean and standard deviation structure using 3 weeks for forecasting.

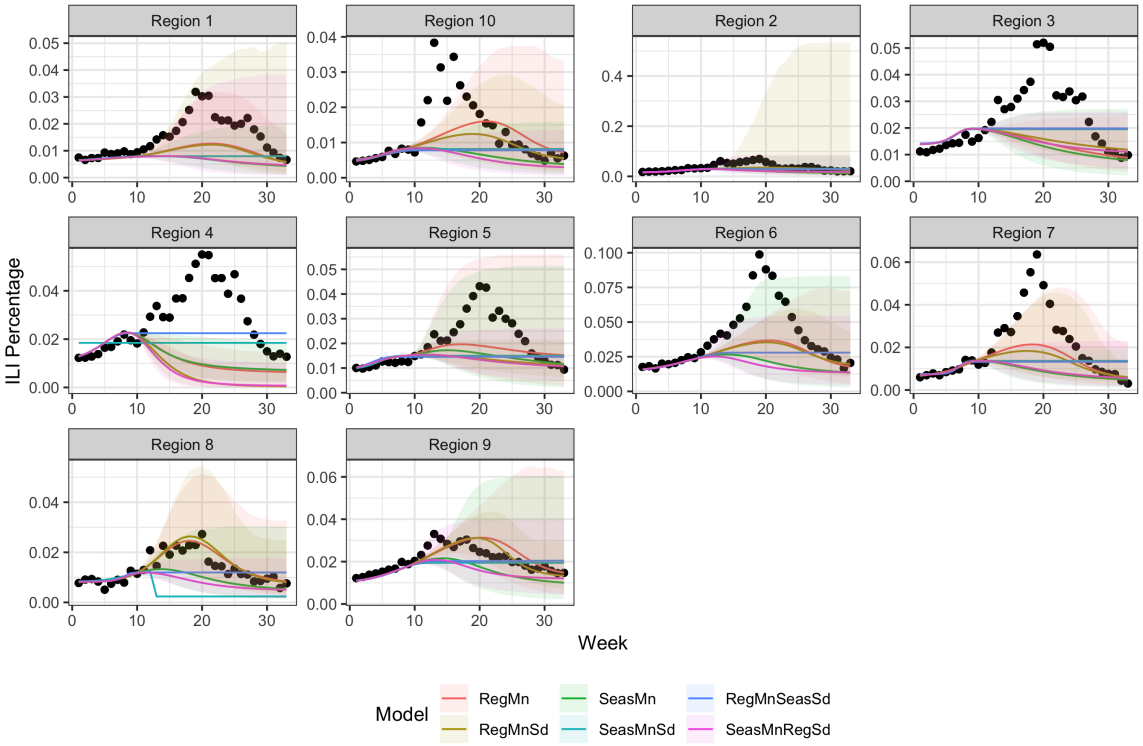

Figure 2.15 Long-term forecasts of all hierarchical structures excluding the independent model for all regions in the 2016-2017 influenza season including only 10 weeks of data from the forecasted season.

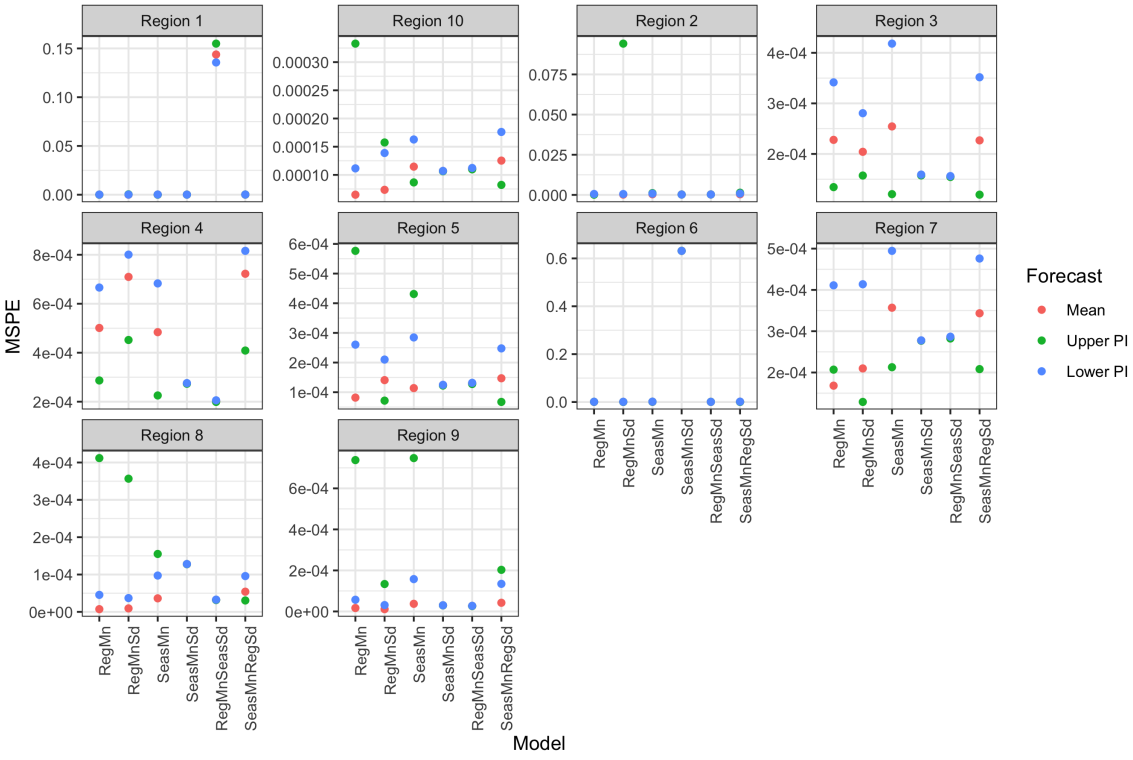

Figure 2.16 Mean square forecast error (MSFE) for the mean forecast and its 95% credible intervals for all models.

CHAPTER 3. BAYESIAN HIERARCHICAL FUNCTIONAL DATA ANALYSIS

3.1 Abstract

Influenza is an illness which affects many people every year. In the past few years, we have seen the great impact influenza can have on the population. The Centers for Disease Control and Prevention (CDC) has created the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) to help them study influenza. In this paper, we use Bayesian hierarchical model with ILINet data to study and forecast influenza. We used Bayesian functional principal component analysis coupled with multiple shrinkage distributions and hierarchical structures. We found that the hierarchical shrinkage distribution had no convergence issues and fit the data well. A season hierarchical structure best fit the past data though a region hierarchical structure preformed the best when forecasting.

3.2 Introduction

Time dependency naturally occurs in many settings: economics, finance, meteorology, longitudinal studies, etc... Typical modeling choices for time series data include ARIMA models (Brockwell and Davis, 2016), dynamic linear models (West and Harrison, 1997), nonlinear models (Priestley, 1978), hidden Markov models (MacDonald and Zucchini, 1997), etc... Time dependent data can also be thought of as snapshots of a functional data process. Consider the influenza season; it runs from early November through late April. The Centers for Disease Control and Prevention (CDC) provides weekly observations of the influenza rate throughout the season, and these observations can be thought of as snapshots of the seasonal influenza rate function over time. Other examples of functional data are yearly birth rates among women in Australia (Hyndman and Shang, 2009), yearly age-specific

female French mortality rates (Hyndman and Shang, 2009), yearly population data (Shang et al., 2013), heart studies where observations are taken at regular time periods (Crainiceanu and Goldsmith, 2010), etc.

When functional data is seasonal, it is natural to wonder if the function will behave similarly in the future. For influenza, the functional data looks similar enough from season to season that we wondered if past seasons could be used to forecast future seasons. With similar thinking, the CDC started the *Predict the Influenza Season Challenge* in November of 2013. This challenge gave the CDC the opportunity to gather and evaluate the methods required to forecast influenza (Centers for Disease Control and Prevention, 2019). The competition requires forecasters to predict the timing, peak, and intensity of the upcoming influenza season. Forecasters could use data from the CDC and any public data available. The competition specifically asked to predict peak week; peak percentage; 1,2,3,4 week ahead forecasts; and influenza onset. The week ahead predictions are considered short-term forecasts whereas the peak week, peak percentage and onset forecasts are long-term forecasts.

Current solutions provide good short-term forecasts (Viboud et al., 2006; Dugas et al., 2013; Paul et al., 2014; Chowell et al., 2016) or long-term forecasts (Nsoesie et al., 2013; Yu et al., 2013b; Xu et al., 2017) but solutions that focus on both are lacking. In addition, some solutions fail to incorporate uncertainty into their forecasts (Dugas et al., 2013; Paul et al., 2014). One good attempt at forecasting influenza is presented by Osthus et al. (2019). In thier paper, they examine the descrepency of mechanistic models and the observed data by modeling the bias. Functional data analysis provides the methodology that will allow us to incorporate our uncertainty and provide accurate short-term and long-term forecasts. Previous uses of functional data analysis to model and predict influenza focused on comparison to time-series models (Oviedo de la Fuente et al., 2018). Crainiceanu and Goldsmith (2010) apply a Bayesian framework to a functional principal component data model and gain uncertainty measures through posteriors. They still face the classical problem in functional

principal component data analysis of selecting the number of basis functions. In this chapter, we present a complete Bayesian functional principal component model using shrinkage distributions to tackle the basis selection problem with a data driven approach.

3.3 ILINet

To study and forecast influenza, we rely on data from the CDC. They provide data from people showing influenza-like illness (ILI). The CDC has formally defined ILI as "fever (temperature of $100°F[37.8°C]$ or greater) and a cough and/or a sore throat without a known cause other than influenza" (CDC, 2017). To collect this data, the CDC has created the United States of America (USA) Outpatient Influenza-like Illness Surveillance Network (ILINet); a network of 2800 outpatient healthcare providers throughout the USA and its territories. The providers report the weekly number of patients they see in their office with ILI and the total number of patients seen that week regardless of the reason for coming in. The weekly data is aggregated into regions by summing all patients with ILI within that region and summing all patients seen within that region. Figure 3.1 shows the 10 regions of the USA arranged by the CDC.

Seasons spanning $2006 - 2014$ were used, excluding $2008 - 2010$ due to H1N1. The influenza season is defined by the CDC as spanning Morbidity and Mortality Weekly Report (MMWR) weeks 40-20 roughly from November through early May. These weeks are the focus since they are primarily when the influenza rate changes the most. For ease of plotting and interpretation, the weeks have relabeled; week 40 is relabeled to week 1 and week 20 is relabeled either week 32 or 33 depending on whether there were 52 or 53 MMWR weeks in the year. In this chapter, we used ILI percentage which is calculated by dividing the raw number of patients with ILI by the number of total patients seen.

Figure 3.2 shows the weekly percentage of patients with ILI plotted against the weeks of the influenza season faceted by season and regions. The shape of the data looks unimodal and is consistent across all regions in all seasons. The peaks of the data typically occur

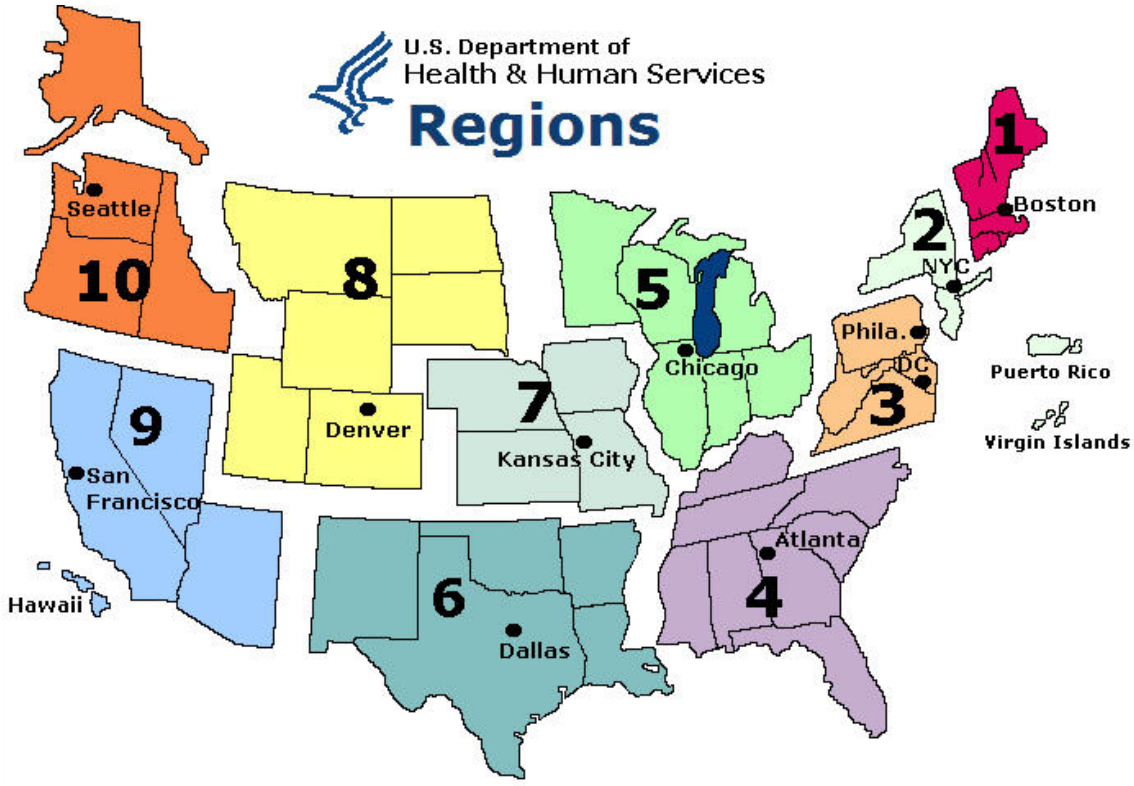

Figure 3.1 The 10 regions of the United States as decided by the CDC (2011).

around weeks 15 through 20, but in some cases, the peak occurs much later as in region 5, season 15-16 where the peak occurs closer to week 25. In all cases, the peak usually corresponds with the times normally associated with high influenza: the cold, wet, winter season. This is consistent with literature showing that high humidity can have an impact on the influenza season (Hemmes et al., 1960; Shaman and Kohn, 2009). There is a seasonal pattern in the data, especially in the timing of the peak. In season 14-15, the peaks all come fairly early around week 12 whereas in season 15-16, the peak weeks are much later, around week 22 or 23. The regional patterns also show up in the data. Region 6 consistently has a higher peak then most of the other regions though the peak percentage looks seasonal.

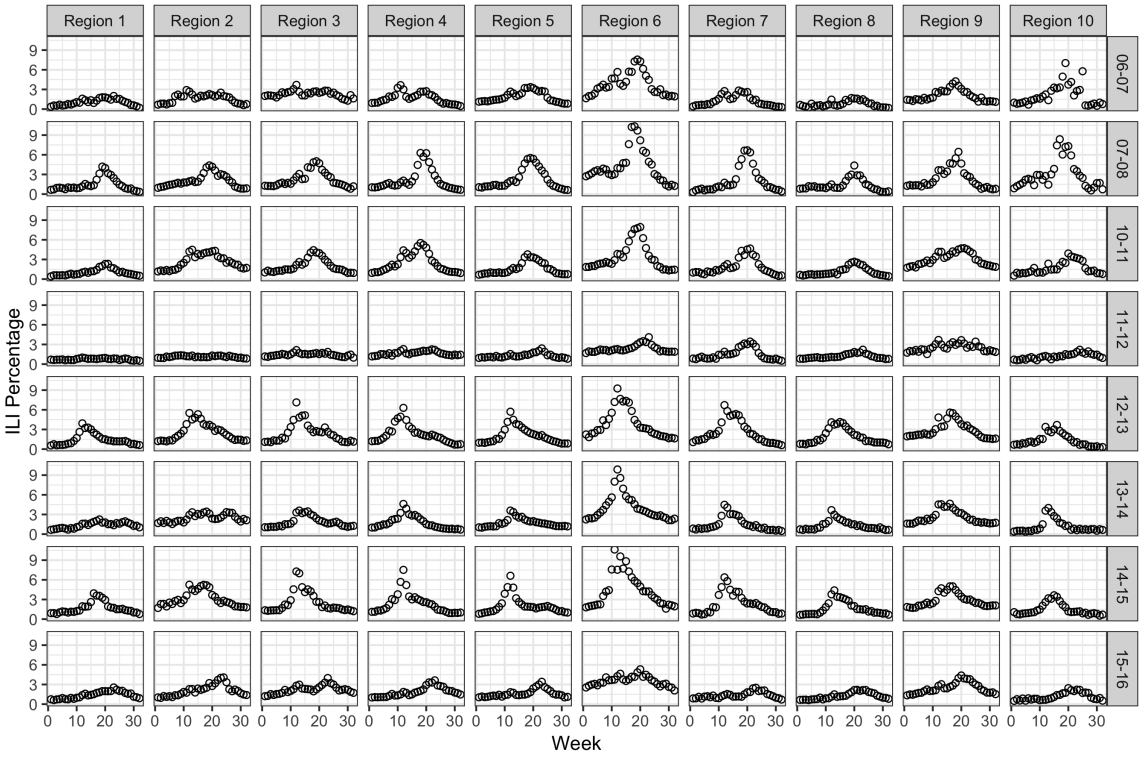

Figure 3.2 Weekly ILI percentage for all regions and seasons faceted by seasons in the rows and regions in the columns.

3.3.1 Registration

Data registration is the act of creating a uniform dataset. For ILINet data, data registration is the act of making season lengths the same and centering the data. ILINet data can have either 32 or 33 weeks in a season depending on how many MMWR weeks are in that year. Functional principal components analysis needs the data to be of the same length so it is necessary to make the seasons a uniform length. This is accomplished by determining which MMWR week Christmas falls on each year and then reassigning that week to a common week across seasons. For example, Christmas in season 14-15 falls on week 12, but in season 15-16, it falls on week 11. In order to reconcile the difference we relabel week 11 in season 15-16 to week 12. This leaves season 15-16 with one more observation than

season 14-15 at the beginning and one more observation for season 14-15 than season 15-16 at the end. Because of the ILI percentage in these weeks is negligible, we simply remove the last observation from season 14-15 and the first observation from 15-16. Now these two season are of the same length and week 12 refers to the same calendar time. Creating equal length seasons means losing data, but the beginning and end of the season do not provide much insight for key forecast points: peak week, peak percentage, ramp up and cool down. For this reason, we assume it is fine to drop these points. At the end of this process, each season was standardized to have 32 weeks. The second facet of data registration is centering the data by subtracting the weekly means from each observation. Equation 3.1 shows the formulas to calculate the centered data, $y^*_{r,s}(w)$ where w is the week of the observation, r is the region, s is the season, R is the total number of regions (10), and S is the total number of seasons (8).

$$y^*_{r,s}(w) = y_{r,s}(w) - \widehat{\mu}(w) \tag{3.1}$$
$$\widehat{\mu}(w) = \frac{1}{RS}\sum_{r=1}^{R}\sum_{s=1}^{S} y_{r,s}(w)$$

To calculate the centered data, first the weekly mean ILI percentage, $\widehat{\mu}(w)$, is calculated across regions and seasons. Then each observation, $y_{r,s}(w)$, get its weekly mean, $\widehat{\mu}(w)$, subtracted from itself. Figure 3.3, shows the data for region 1 in season 10-11 after the length has been corrected. The different shapes correspond the data before centering and after centering. Since most of the seasons have the peak occuring early in the season, the weekly means for weeks 10-15 are large realtive to season 10-10. This accounts for the change in the peak.

3.4 Methodology

In this section, we will describe function principal component analysis. Next, we propose multiple shrinkage distributions for the coefficients and hierarchical models for these

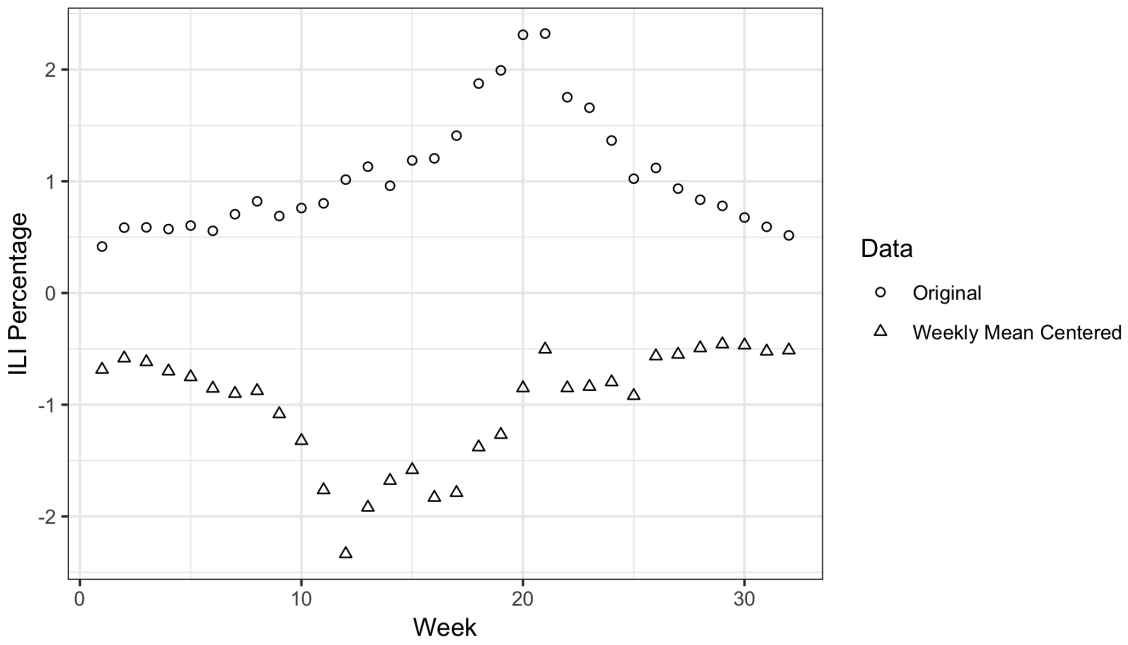

Figure 3.3 Data for region 1 in season 10-11 after the length has been adjusted. The shapes correspond to the data before and after centering.

coefficients. Then the estimation details used in this analysis will be presented and lastly, the methods used to assess model fit will be reviewed.

3.4.1 Data Model

Let $y_{r,s}(w)$ be the registered ILI percentages for week w in region r and season s. Equation 3.2 is a functional data model where $y_{r,s}(w)$ is the composition of the true underlying smooth function, $\psi_{r,s}(w)$, with some observation error, $\epsilon_{r,s}(w)$.

$$y_{r,s}(w) = \psi_{r,s}(w) + \epsilon_{r,s}(w) \quad \epsilon_{r,s}(w) \overset{ind}{\sim} N(0, \sigma_\epsilon^2) \tag{3.2}$$

Functional principled component analysis (FPCA) is a method developed by Dauxois and Pousse (1976, 1982) to model $\psi_{r,s}(w)$. Equation 3.3 shows a typical representation of FPCA where $\mu(w)$ is the function of weekly means and K is the number of basis functions.

The goal for FPCA is to decompose an infinite dimension functional space to into a finite set of principal components (Ramsay and Silverman, 2005; Di et al., 2009). To compute the principal components from the data, the data variance-covariance matrix needs to be estimated. Once estimated, it smoothed using a non-parametric smoother (Di et al., 2009). We estimate the variance-covariance matrix using methods method of moments (MoM) via the `cov` function in `R` and smoothed by using the `spm` function from the `SemiPar` package (Wand, 2018) which employs a semiparametric regression model using the mixed model representation of penalized splines (Ruppert et al., 2003). Finally, the eigenvectors, v_k, are derived from the smoothed empirical variance-covariance matrix and used as the principal components for the data (Di et al., 2009).

$$\psi_{r,s}(w) = \mu(w) + \sum_{k=1}^{K} \beta_{r,s,k} \mathrm{v}_k(w) \tag{3.3}$$

Wang et al. (2015) points out in their review of functional data analysis that one of the biggest issues with functional principal component analysis is determining how many principal components should be used i.e. how large k should be? A common solution is to consider the fraction of variance explained by the principal components. A cutoff value is chosen and once that amount of variance is accounted for, that corresponding K is the number of principal components used. Other solutions use AIC and BIC as well as leave one out cross validation (LOO CV) but they tend to overfit the data by including too many components (Wang et al., 2015). Our solution to this problem is to apply hierarchical shrinkage distributions to the β parameters and let the data decide how many components to use.

3.4.2 Shrinkage Distributions

In this subsection, we will introduce several shrinkage distributions. Any of these shrinkage distributions can be used in the hierarchical structures we will propose.

3.4.2.1 Horseshoe Distribution

The horseshoe prior was first proposed in *Handling Sparsity via the Horseshoe* by Carvalho et al. (2009). This paper provided the framework for a fully Bayesian approach to sparse supervised-learning. The problem setup for this distribution is a regression problem where there are many predictors and we want to know which are relevant to the response. This distribution will allow the data to suggest which predictors are meaningful at capturing the data. The horseshoe distribution is noted in equation 3.4 where HS is a scale mixture of normals defined in equation 3.5. It is important to note that there is no closed form marginal distribution of β.

$$\beta_{r,s,k} \overset{ind}{\sim} \text{HS}(\tau_{r,s}) \tag{3.4}$$

$$\begin{aligned} \beta_{r,s,k}|\lambda_{r,s,k},\tau_{r,s} &\overset{ind}{\sim} N(0,\lambda_{r,s,k}^2\tau_{r,s}^2) \\ \lambda_{r,s,k} &\overset{ind}{\sim} Ca^{+}(0,1) \end{aligned} \tag{3.5}$$

In equation 3.5, $\lambda_{r,s,k}$ and $\tau_{r,s}$ represent the local and global shrinkage parameters, respectively. These parameters determine which explanatory variables are of consequence to the response. By assigning heavy-tailed half Cauchy distributions to the shrinkage parameters, it allows the data to grow λ to large values greater than the threshold, τ. The Cauchy distribution also allows the data to push the shrinkage parameters towards 0 by having mass at 0. The Cauchy distribution is the key to the horseshoe distribution. A priori, it puts most of the probability on total shrinkage towards 0 or no shrinkage.

If we think about FPCA model as a regression problem where the principal components are the explanatory variables, we can apply the horseshoe distribution (or any other shrinkage distribution) to the coefficients on the principal components. A FPCA model using the

horseshoe distribution is written out in equation 3.6.

$$
\begin{aligned}
y_{r,s}(w) - \widehat{\mu}(w) &= \sum_{k=1}^{K} \beta_{r,s,k} \mathrm{v}_k(w) + \epsilon_{r,s}(w) \\
\epsilon_{r,s}(w) &\overset{ind}{\sim} N(0, \sigma_\epsilon^2) \\
\beta_{r,s,k} &\overset{ind}{\sim} N(0, \lambda_{r,s,k}^2 \tau_{r,s}^2) \\
\lambda_{r,s,k} &\overset{ind}{\sim} Ca^+(0,1)
\end{aligned} \tag{3.6}
$$

In equation 3.6, $\epsilon_{r,s}(w)$, $\beta_{r,s,k}$, and $\lambda_{r,s,k}$ are independent of each other, and $Ca^+(a,b)$ refers to a positive half-Cauchy distribution.

While the horseshoe distribution is a well known shrinkage distribution, there have been noted convergences issues with the MCMC (Piironen and Vehtari, 2017). Thankfully, there are other shrinkage distributions that can be used.

3.4.2.2 Regularized Horseshoe

Piironen and Vehtari (2017) mention the potential pitfalls of the horseshoe distribution and propose a solution which they have named the regularized horseshoe. The regularized horseshoe distribution behaves similarly to the horseshoe distribution but allows one to specify a minimal level of regularization to the large values/signals. With the horseshoe distribution, there is no regularization on the large values of λ. This can leave the β posteriors diffuse to extremely large values which can cause nonidentification or weak identification of the likelihood (Betancourt, 2018). The regularized horseshoe prior is written in equation 3.7 where RHS is defined in equation 3.8.

$$
\beta_{r,s,k} \overset{ind}{\sim} \mathrm{RHS}(\tau_{r,s}) \tag{3.7}
$$

$$\begin{aligned}
\beta_{r,s,k}|\lambda_{r,s,k},\tau_{r,s} &\sim N(0,\tilde{\lambda}^2_{r,s,k}\tau^2_{r,s}) \\
\tilde{\lambda}_{r,s,k} &= \frac{\upsilon^2_{r,s}\lambda^2_{r,s,k}}{\upsilon^2_{r,s}+\lambda^2_{r,s,k}\tau^2_{r,s}} \\
\lambda_{r,s,k} &\overset{ind}{\sim} Ca^+(0,1) \\
\upsilon^2_{r,s} &\sim \text{Inv-Gamma}(a,b)
\end{aligned} \tag{3.8}$$

The υ parameter is what provides the regularization to the λ parameters. When the parameters for the distribution on υ (a, b) are chosen carefully, it ensures that the coefficient posteriors will be contained around 0.

3.4.2.3 Hierarchical Shrinkage

Another alternative shrinkage distribution is a simple twist on the horseshoe. Instead of placing a half-Cauchy distribution on the λ parameters, a standard half-t is used shown in equation 3.9. Using the half-t distribution provides improved sampling for the β parameters. Piironen and Vehtari (2017) list this as a suggestion, but claim their regularized horseshoe (3.7) outperforms this distribution. One potential drawback is that the distribution is less sparsifying due the thinner tails of the t distribution though in this project we did not find this to be true.

$$\begin{aligned}
\beta_{r,s,k} &\overset{ind}{\sim} N(0,\lambda^2_{r,s,k}\tau^2_{r,s}) \\
\lambda_{r,s,k} &\overset{ind}{\sim} t^+_v(0,1)
\end{aligned} \tag{3.9}$$

3.4.2.4 Bayesian Lasso

One more possible shrinkage distribution is the Bayesian LASSO (equation 3.10) which places independent Laplace (double-exponential) distributions directly on the β parameters (Park and Casella, 2008; Hans, 2009). It is the Bayesian counterpart to the frequentist

LASSO (Tibshirani, 1996) whose estimates are L_1-penalized least squares estimates. Another possible shrinkage distribution similar to Bayesian LASSO is Bayesian ridge regression though we did not include it here.

$$\beta_{r,s,k} \overset{ind}{\sim} \text{Laplace}(0, \tau_{r,s}) \tag{3.10}$$
$$\text{Laplace}(x; 0, \tau_{r,s}) = \frac{1}{2\tau_{r,s}} \exp(-\frac{|x|}{\tau_{r,s}})$$

3.4.3 Hierarchy Structures

Research shows that a model can benefit from borrowing information through a hierarchical structure (Mugglin et al., 2002; Michaud, 2016; Yu et al., 2013b). One benefit is posterior shrinkage towards the mean. A natural hierarchical structure for our data could borrow information across regions or seasons. One way of accomplishing this is to have similar shrinkage patterns across regions or seasons. In equation 3.6, we have set up models that borrow information within each region and season. Each region-season combination is conditionally independent as in equation 3.11 where $\pi_1(\theta_0)$ and $\pi_2(\theta_0)$ are the respective distribution and prior on the shrinkage parameters. This gives each region-season combination the flexibility to choose its own principal components that are relevant to that particular combination.

$$\tau_{r,s} \overset{ind}{\sim} \pi_1(\theta_0) \tag{3.11}$$
$$\lambda_{r,s,k} \overset{ind}{\sim} \pi_2(\theta_0)$$

But if we wanted to have a similar shrinkage pattern within a region across the different seasons, we could replace $\lambda_{r,s,k}$ and $\tau_{r,s}$ with $\lambda_{r,k}$ and τ_r, respectively and set up our shrinkage parameters as in equation 3.12.

$$\tau_r \overset{ind}{\sim} \pi_1(\theta_0) \tag{3.12}$$
$$\lambda_{r,k} \overset{ind}{\sim} \pi_2(\theta_0)$$

(3.13)

Similarly, we could want the same shrinkage pattern within a season across all regions and replace $\lambda_{r,s,k}$ and $\tau_{r,s}$ with $\lambda_{s,k}$ and τ_s, respectively as in equation 3.14.

$$\tau_s \overset{ind}{\sim} \pi_1(\theta_0) \tag{3.14}$$

$$\lambda_{s,k} \overset{ind}{\sim} \pi_2(\theta_0) \tag{3.15}$$

3.4.4 Basis Choices

Within our modeling choices is the need to choose basis functions. We chose to go with principal components, but there are many possible basis choices such as: Fourier, polynomial, etc. We generated thirty principal components. In the literature we reviewed, the highest number of principal components ever used was ten, but in this project we wanted to show that given more principal components than needed, the model can pick out the relevant ones. Figure 3.4 shows the thirty principal components created from the data. The first nine or ten seem the most general and are smoother than the others while the rest get increasingly more specialized, capturing specific features of the data noise.

3.4.5 Forecasting

One of the motivations for considering FPCA was to create reasonable forecasts that provided good mean forecasts and reasonable uncertainty. There are other methods to forecast with FPCA (Hyndman and Shang, 2009; Shang et al., 2013) though they lack a Bayesian approach and a cohesive method to choose the number of basis functions. Our model has both and lends itself to forecasting very easily. The forecasts are generated by including a certain number of weeks of the forecasted season and estimating the β parameters. The β posterior samples are plugged into Equation 3.3 and then plugged into Equation 3.2 to obtain forecasts for the rest of the season. To assess whether or not the shrinkage distributions and the different hierarchical structures are good at forecasting, both short and long-term forecasts were analyzed. For long term forecasts, one can keep

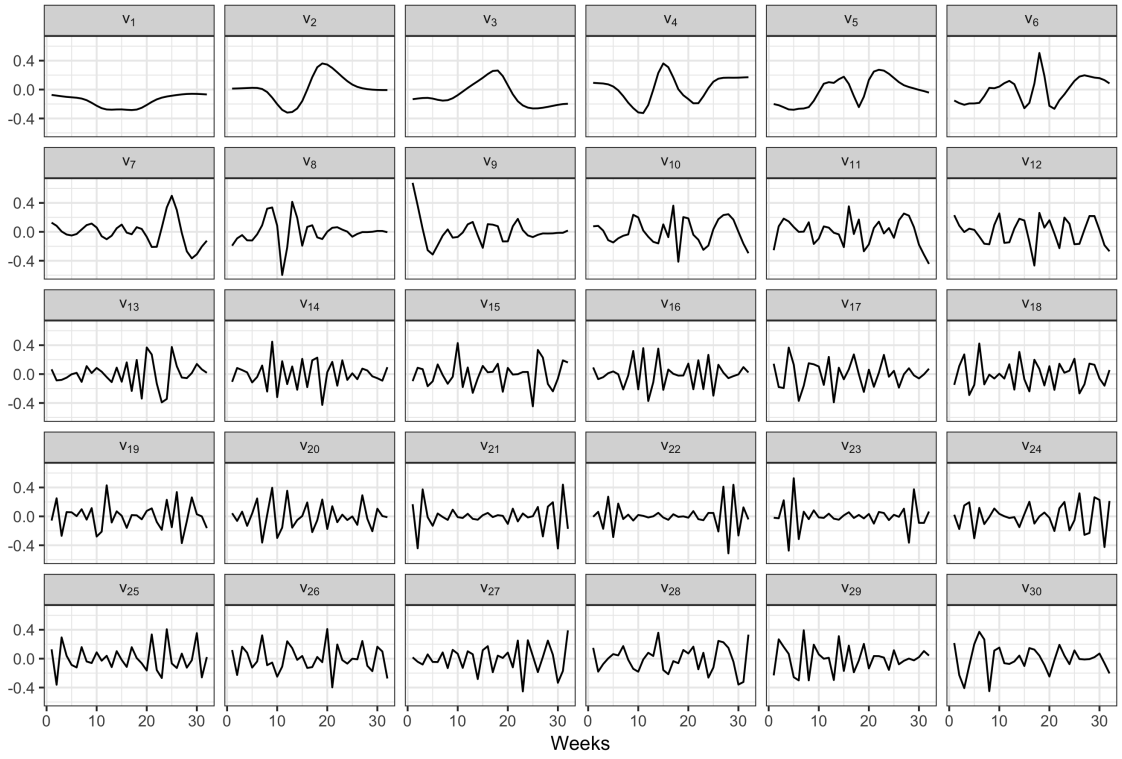

Figure 3.4 Thirty created eignenfunctions from the smoothed empirical variance-covariance matrix estimated from the registered data.

the forecasts as is and look at what will happen for the rest of the season or you can take your forecasts and see what it predicts 1, 2, 3, or 4 weeks out for short term forecasts. This process for generating forecasts is the same regardless of the hierarchical structure or shrinkage prior.

3.4.6 Model Checking

In this subsection, we will describe how we fit the models and how we are going to assess them. Fist, we lay out the details needed to fit the models and how the parameters were estimated using Markov chain Monte Carlo (MCMC). Next, we describe how to assess convergence of the MCMC. Lastly, the methods used to assess model fit and forecasting fit

are described. These statistics will help us discern which hierarchical structure fits the data best. We also discuss how to compare the shrinkage properties of the distributions.

3.4.6.1 Estimation

In order for most of the shrinkage distributions to be complete, there needs to be a prior on τ. In equation 3.16, the priors on τ and σ are presented. As in Carvalho et al. (2009), we used a half-Cauchy prior on τ. And for the data standard deviation, σ, we also assign a half-Cauchy prior. These priors are standard non-informative priors for standard deviation parameters (Gelman, 2006b; Polson and Scott, 2012).

$$\begin{aligned} \tau_{r,s}, \tau_r, \tau_s &\overset{ind}{\sim} Ca^+(0,1) \\ \sigma &\overset{ind}{\sim} Ca^+(0,1) \end{aligned} \tag{3.16}$$

We can also included an independent prior on the β parameters using no hierarchical structure or sparsity inducing distribution for comparison in equation 3.17. Each regression coefficient is assigned an independent vague normal prior.

$$\beta_{r,s,k} \overset{ind}{\sim} N(0, 10) \tag{3.17}$$

The models were fit using Hamiltonian Monte Carlo via `Stan` (Stan Development Team, 2016) through `R` (R Core Team, 2016). One chain ran for 6000 iterations with half of that used for burn-in. Random starting points were used and generated by `STAN`.

3.4.6.2 Convergence Check

To assess issues with convergence, we used Geweke's diagnostic. Geweke (1992) created a convergence diagnostic for Markov chains in which if there are no causes for concern, the Geweke statistic should look like draws from a standard normal distribution. Trace plots were also used in a visual inspection for issues with convergence.

3.4.6.3 Model Fit

Three hierarchical structures have been proposed and multiple shrinkage distributions have been suggested, but which model fits the best? We will compare root mean squared error (RMSE) and root mean square forecast error (RMSFE) to answer this question. The formulas both are listed in equation 3.18. In the formulas, $y_{r,s}(w)$ is the data from region r, season s and week w; $\widehat{\tau_{r,s}}(w)$ is the estimated smooth underlying function; and $y^*_{r,s}(w)$ and $\widehat{\tau^*_{r,s}}(w)$ are their counterparts in the forecasted season.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{r,s}(w) - \widehat{\tau_{r,s}}(w)\right)^2} \tag{3.18}$$
$$RMSFE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y^*_{r,s}(w) - \widehat{\tau^*_{r,s}}(w)\right)^2}$$

RMSE is the square root of average squared errors between what is predicted and the actual observations. This allows us to measure how close our fit is to observed ILI rates. We used the RMSE instead of the MSE so it can be interpreted on the data scale. RMSFE is similar except the it uses data from the forecasted season i.e. $y^*_{r,s}(w)$ is the ILI percentages from the forecasted season and $\widehat{\tau^*_{r,s}}(w)$ is the forecasted smooth mean.

3.4.7 M_{eff}

To compare shrinkage abilities between shrinkage distributions, we need to be able to quantify the amount of shrinkage. One option would be to look at the posterior probability of the beta parameters being within some neighborhood of 0. This transfers the burden to deciding on an appropriate neighborhood. Consider the interval of $(-1, 1)$; this could be small enough or it could be too big. It would would really depend on the individual problem. Piironen and Vehtari (2017) present an estimator, M_{eff}, which estimates the effective number of nonzero coefficients i.e. the number of coefficients which have not been shrunk towards 0. M_{eff} is presented in equation 3.19 where $s_k^2 = Var(\mathrm{v}_k)$.

$$\mathrm{M}_{eff} = \sum_{k=1}^{D}(1-\kappa_k) \tag{3.19}$$

$$E(\kappa_k|\tau,\sigma) = \frac{1}{1+n\tau^2 s_k^2 \lambda_k^2/\sigma^2}$$

By plugging in posterior samples for τ, λ and σ, we can get posterior estimates of M_{eff} and compare the shrinkage properties of each distribution.

3.5 Analysis of ILINet

3.5.1 Convergence Check

Figure 3.5 plots the Geweke diagnostic qqplot for all parameters in the different models. The Geweke diagnostic should look like random draws from a standard normal distribution. For most models, the values in Figure 3.5 follow the straight line and show no cause for concern, but the regularized horseshoe and the horseshoe prior (left plot) do show serious cause for concern.

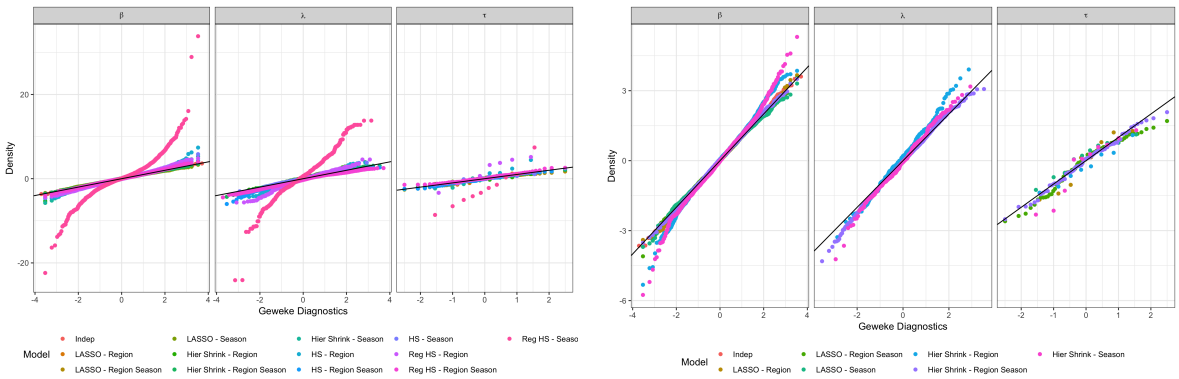

Figure 3.5 Geweke diagnostics Q-Q plot for the parameters of the different models including the horseshoe prior and regularized horseshoe prior on the left and excluding those priors on the right.

3.5.2 Basis Selection

Figure 3.6 shows the posterior mean estimates and 95% credible intervals of the β parameters for the first ten principal components for Regions 7 & 9. Regions 7 covers the Midwest including Iowa and region 9 covers the west coast including California; these two locations have personal meaning to me but also provide a view of results from two distinct areas of the country. Most of the mean estimates of β and their credible intervals hover around 0. Only a few estimates are away from 0 (usually the first or second). As we hoped, the data is able to inform that only 2 or 3 principal components are being used to model the data and the rest are not contributing much.

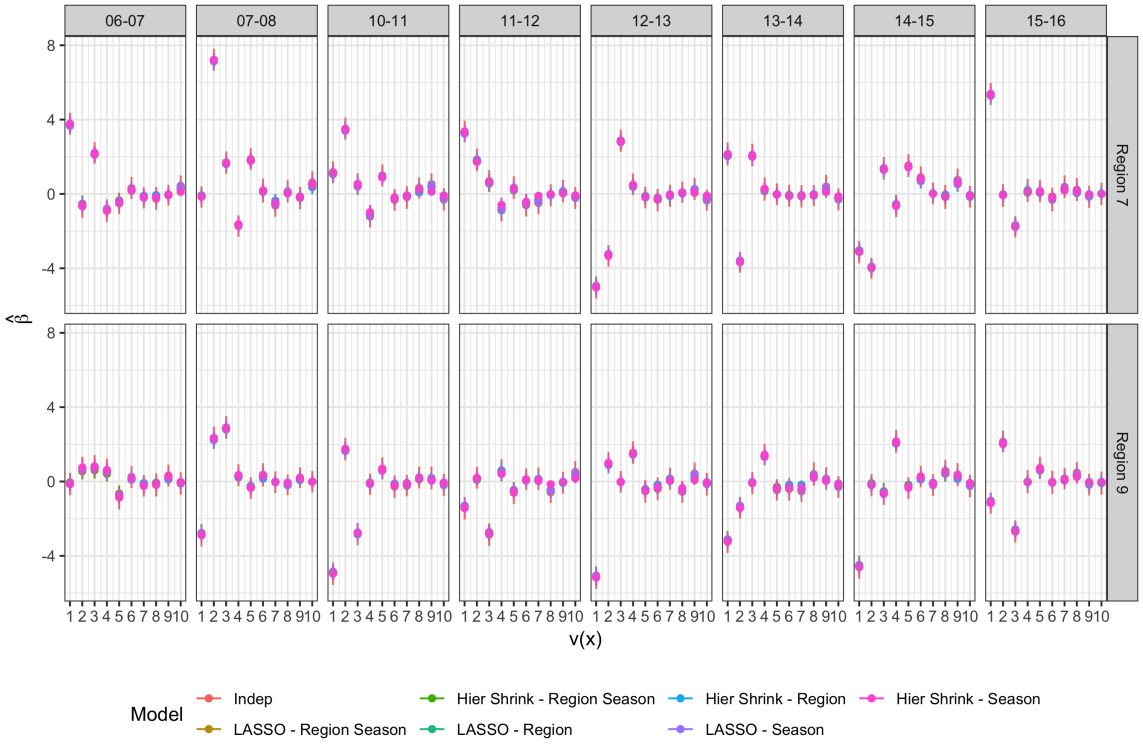

Figure 3.6 Posterior mean estimates and 95% credible intervals of the β parameters for the first ten principal components for Regions 7 & 9.

Figure 3.7 shows the posterior mean estimates and 95% credible intervals of the β parameters for the first ten principal components for $10-11$ and $11-12$ seasons.

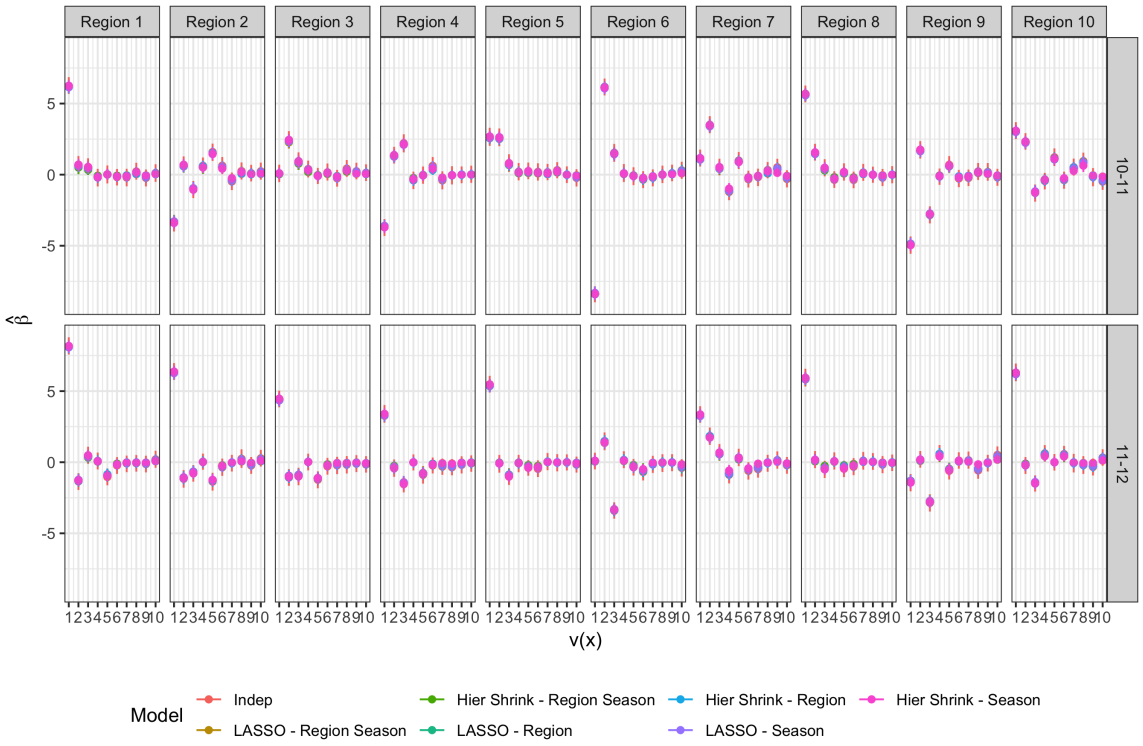

Figure 3.7 Posterior mean estimates and 95% credible intervals of the β parameters for the first six principal components for Season $10-11$ and $11-12$.

An interesting pattern from Figures 3.6 and 3.7 is that there is more of a regional pattern in the β estimates than there is a seasonal pattern. There is more consistency in which β parameters are away from zero across seasons within a region than there is across regions within a season.

3.5.3 Model Fit

Figure 3.8 shows the posterior mean fit with their 95% credible intervals on the 10-11 season data for all models. This season has a good range of peak percentages across the regions so it is a good year to analyze model fits. The independent models and the region-

season models have larger credible intervals than the rest, but all models have similar mean fits. One interesting aspect of the FPCA model is that it allows for more flexibility; it is not limited to keeping one peak or specific shape. It allows for dips and rises along the way to and from the overall peak.

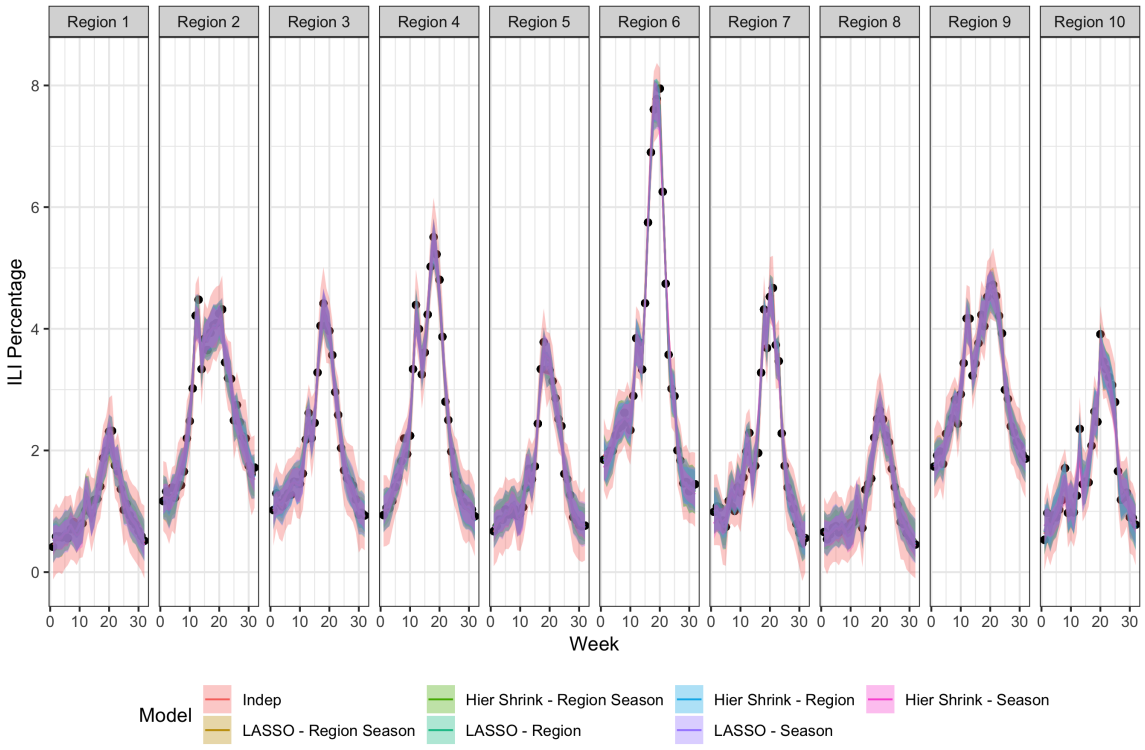

Figure 3.8 Posterior mean fit with 95% credible intervals for the 10-11 season for all models.

Beyond visually diagnosing the fit of the model, we can estimate the error of the model fits using RSME. Table 3.1 shows the root mean square error and average 95% credible interval width for each model with the smallest of each subtracted from the others. The independent model has the lowest RMSE but it also has the largest average 95% credible interval width. This is to be expected since the hierarchical structures are known to be useful in providing tighter credible intervals. The horseshoe distribution and regularized horseshoe distribution had the lowest credible interval width and low RMSE but with their

convergence issues we cannot trust these results. The hierarchical shrinkage prior shows similar average credible interval widths and RSME to these distribution but did not show issues with convergence.

Table 3.1 This table shows the root mean square error and average 95% credible interval width of all models with the smallest of each subtracted from the others.

Model	Δ RMSE	Δ Avg CI Width
Indep	0.000	0.735
LASSO - Region Season	0.018	0.303
Hier Shrink - Region Season	0.033	0.211
LASSO - Region	0.017	0.308
Hier Shrink - Region	0.055	0.135
LASSO - Season	0.016	0.316
Hier Shrink - Season	0.059	0.139

3.5.4 $\mathbf{M}_{eff}$

Figure 3.9 shows the posterior samples of M_{eff} in a boxplot for each season and region. The color depicts the overall model and is faceted by hierarchical structure in the rows and regions in the column. The hierarchical structure's effect on the number of parameters that are allowed to be nonzero is strong. Regardless of the shrinkage distribution, in the region hierarchical structure, there is a unique pattern across the seasons within a specific region, and in the season hierarchical structure, there is a season pattern across the regions. Compared to these, the region-season hierarchical structure is much more free. One especially beneficial pattern of including a hierarchical structure across regions or seasons is that the number of nonzero coefficients is always low; M_{eff} hovers around 2 for both whereas with the region-season hierarchical structure, the freedom allows the number to be as high as 6 in some cases.

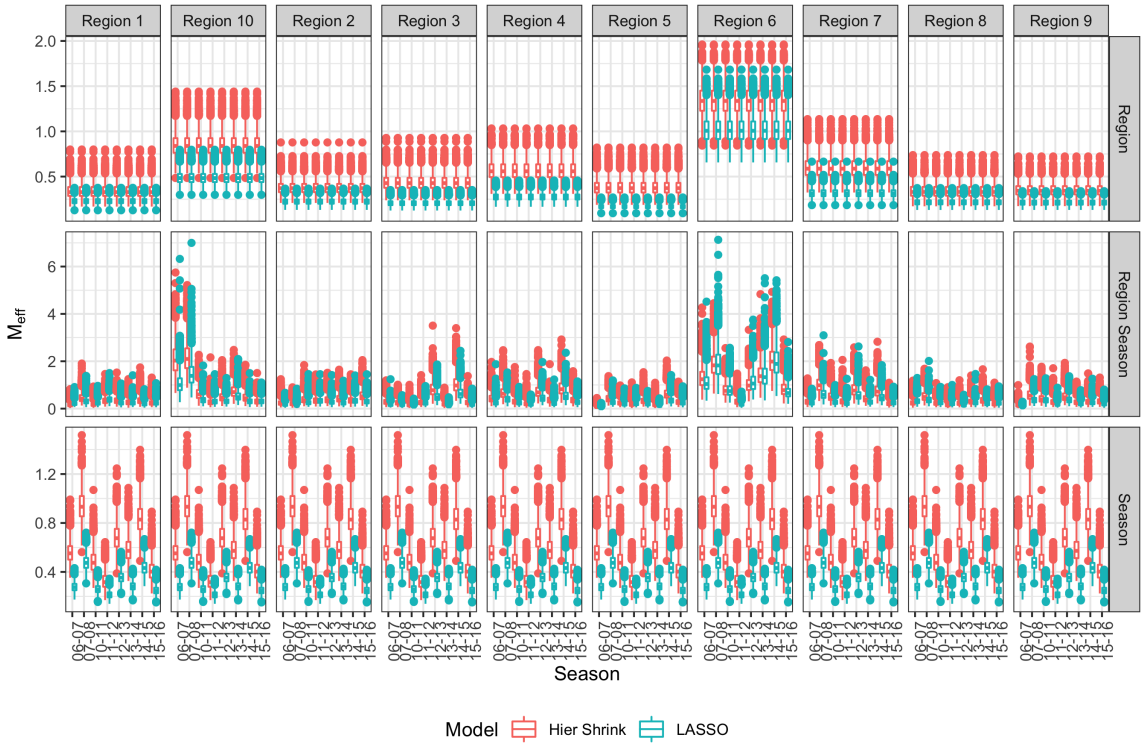

Figure 3.9 Pposterior samples of M_{eff} i.e. the number of β coefficients not shrunk to zero faceted by regions in the columns and hierarchical structure in the rows. The colors represent the different shrinkage priors.

3.5.5 Forecasting

Figure 3.10 shows the 1, 2, 3, and 4 week ahead forecasts for the whole season and their 95% credible intervals colored by hierarchical structure and faceted by region. The short term forecasts show what is predicted to happen in the immediate future. The 1 week ahead forecasts perform well in that the forecasted percent ILI are fairly close to the reported ILI percentages and the credible intervals are mostly reasonable. We see similarities in the 2 week ahead forecasts. The 3 and 4 week ahead forecast start to see larger credible interval widths and the forecasts are further from the observed ILI.

Figure 3.11 shows the long-term mean forecasts and 95% credible intervals on the original data. The columns represent the number of weeks of the new season included in the

forecast and the rows correspond to the different regions. The colors correspond to the different hierarchical structures using the hierarchical shrinkage prior. We only looked at the hierarchical shrinkage prior because the LASSO prior and independent prior provided too wide of credible bands to be of any practical use in forecasting. The regional, seasonal, and region-season hierarchical structures borrow information to provide much more reasonable forecast intervals. As more data is included for the forecast, the forecasts improve.

With small amounts of data from the forecsted season included, we still see good long term forecasts. In particular, season peak forecasts are close to the true values with little data included in the forecast. Figure 3.12 and 3.13 show the estimated posterior peak week and peak percentage and their 95% credible intervals as a function of the number of weeks included in the forecasted season, respectively. Especially with peak week, with few weeks included, the estimated peak week is fairly close to the observed peak week. Peak percentage needs more time to adjust, but most of the credible bands are covering the true value.

To see if the peaks in the forecasted weeks were unusual, the peak week and peak percentage for the past seasons are plotted in Figures 3.14 and 3.15. For both the peak week and peak percentage, the values are generally in the range of reasonable values but there are some reasons we could see some issues with forecasting. In Region 6, the peak week seems reasonable, but in the recent past there have been some low peaks. These peaks are bringing down the estimated peak percentage.

RMSFE is looked at to assess the raw errors in the forecasts of the different hierarchical structure. Along with the forecasts, we are also interested in if our uncertainty is well represented; the average 95% credible interval width can be used to examine this. Figure 3.16 shows the RMSFE and average 95% credible interval width by the number of weeks used in the forecast. Immediately, we recognize that the independent prior (no hierarchical structure) has a very large average 95% credible interval width relative to the other hierarchical structures which have all converged. This is because without borrowing any information

on which coefficients to shrink, the independent model is very unsure of what the forecast should look like whereas with the different hierarchies, they can borrow information and are less unsure about which coefficients to shrink in the forecasted season.

The RSMFE shows a more interesting pattern. The region hierarchical structure has a smaller error than the other hierarchical structures and its error is consistently lower no matter the number of weeks included in the forecast until the end of the forecast where the season hierarchical structure RMSFE matches it. All hierarchical structures follow the same logical pattern: the RMSFE decreases as the number of included weeks increases. The season hierarchical structure has a dramatic decrease in RMSFE at week 7 and 15. These time points are right around the ramp up and peak of the season. Due to the nature of the season hierarchical structure, borrowing information across regions within a season, its accuracy is going to increase as the season goes on.

3.6 Discussion

In this paper, we focused on methodology that would allow us to model and forecast influenza. Functional data analysis methods are especially useful in forecasting influenza. Principal components were used as the basis functions of the functional space. Shrinkage distributions were used to control the number of principal components included in the model. This allowed for a complete model-based, data-driven approach to this problem which has been handled in an ad-hoc manner by others in the past. Multiple shrinkage distributions were compared and we found that the hierarchical shrinkage distribution provided similar results to the horseshoe distribution while not showing convergence issues. This framework was assessed by its forecast on a new season. Both short-term and long-term forecast preformed well. Forecasting the peak week with very little data of the forecasted season worked especially well and even the forecasted peak percentage was reasonably close. The short-term forecasts performed well too. In all cases, borrowing information across regions

or season through hierarchical structures improved our credible intervals by making them tighter.

In this project, we are not only concerned with having reasonable mean forecasts, but also measures of uncertainty in our forecasts. By using a Bayesian model, we were able to obtain credible intervals for all our forecasts. The independent model fared comparably well to the hierarchical structures when modeling the data, but when forecasting with the independent model, the mean forecasts were worse and the credible intervals were very wide. The intervals were so wide they were unusable. If you simply want to model and perform inference on past seasons, an independent model would suffice; though if interested in any type of forecasting, the independent model should not be used. If interested in forecasting, the hierarchical shrinkage prior with a hierarchical structure on the shrinkage parameters would be good to use. The region hierarchical structure performed better at forecasting the season peak week and peek percentage than the other hierarchical structures.

One area for concern is the aggregation of ILINet data. Regions 2, 9, and 10 include states and/or territories that are not connected to the rest of the region. For example, it would be reasonable to think that Puerto Rico and the Virgin Islands have very little in common with New York and should have their own region, yet they are aggregated into Region 2. Alaska and Hawaii have similar issues with their associated regions. If attempting to conduct a spatial analysis it might affect how the movement of the peak of the influenza season is modeled.

In this paper we focused on using principal components for the basis but there are many other choices. Of particular interest is choosing basis function that promote a smoother overall function such polynomial basis functions. We would also like to comepare our work to Osthus et al. (2019) as they are a leading forecasting model. Work with assessing the quality of our uncertainty is needed. We can look at coverage in our forecast credible intervals to see how reasonable they are in addition to looking at average width. There is also more work to be done on computation time. All of these models took a considerable

amount of time ($\approx$ 5 days). An empirical Bayes approach to these models where we plug in estimates for the shrinkage parameters estimated from the past seasons and forecast the new season might help to decrease computation time.

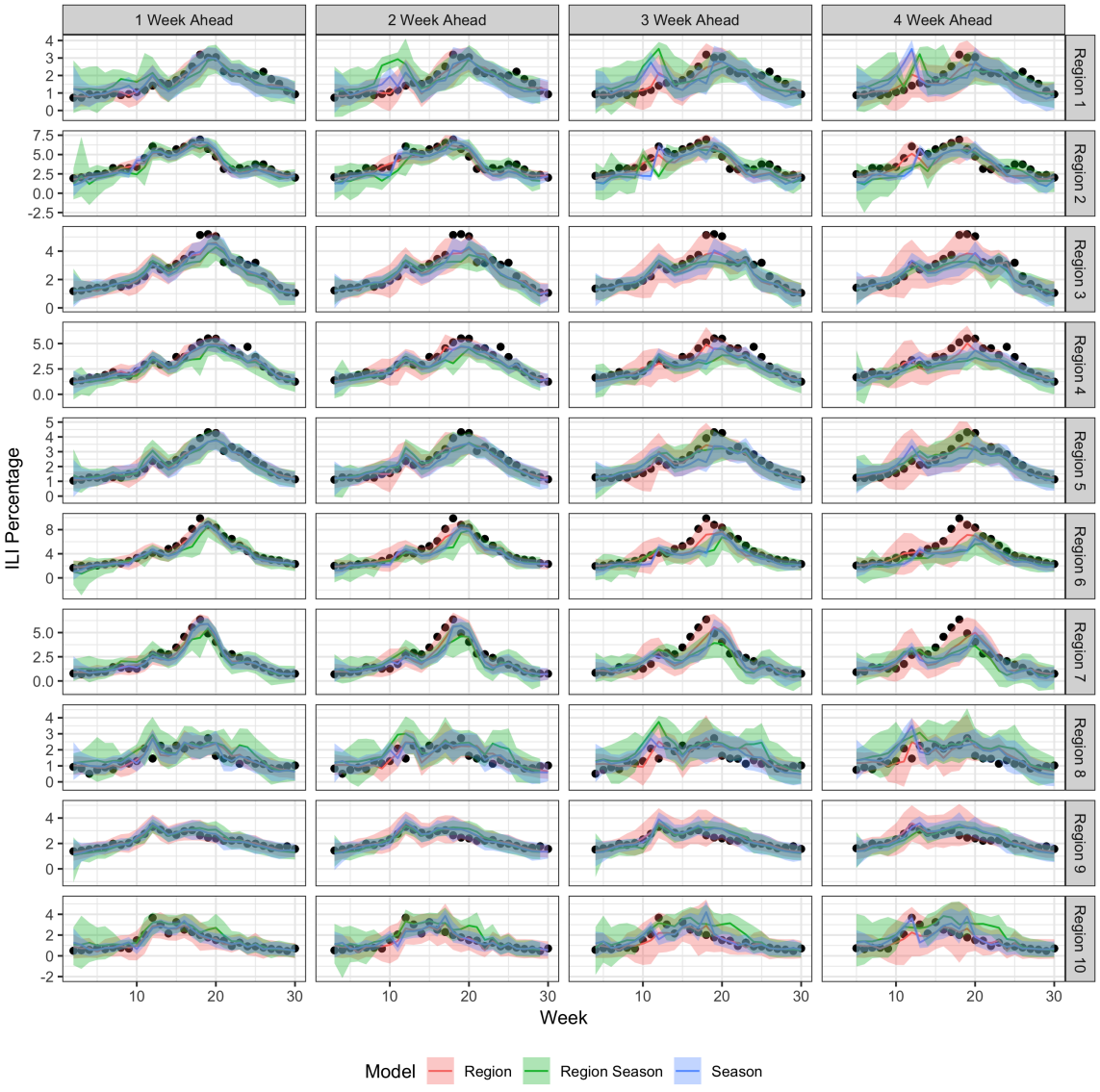

Figure 3.10 1, 2, 3, and 4 week ahead forecasts and 95% credible intervals from the model using the hierarchical shrinkage distribution colored by hierarchical structure and faceted by region.

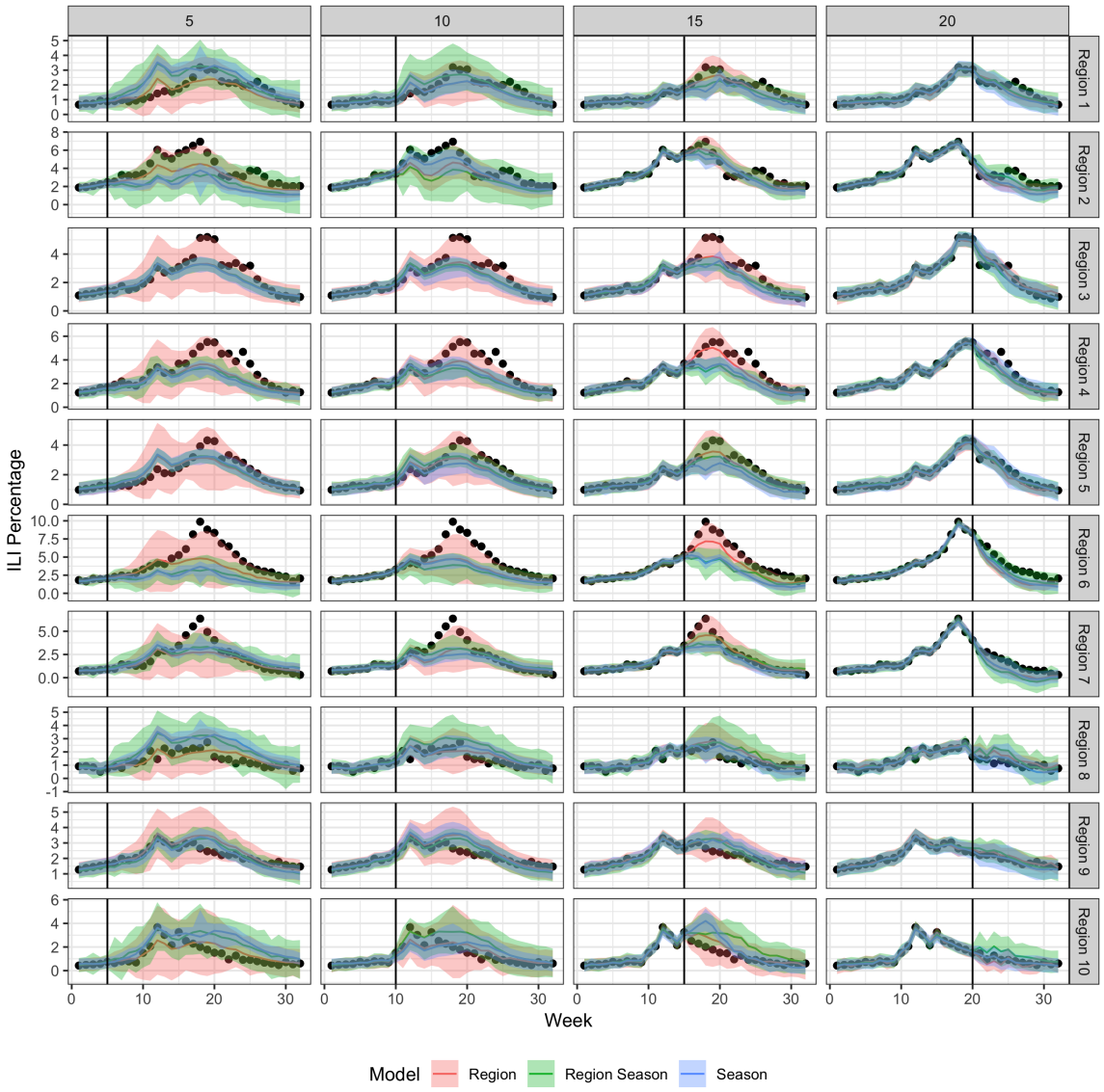

Figure 3.11 Long-term posterior mean forecasts and 95% credible intervals on the original data faceted by region and number of weeks from the forecasted season included in the estimation.

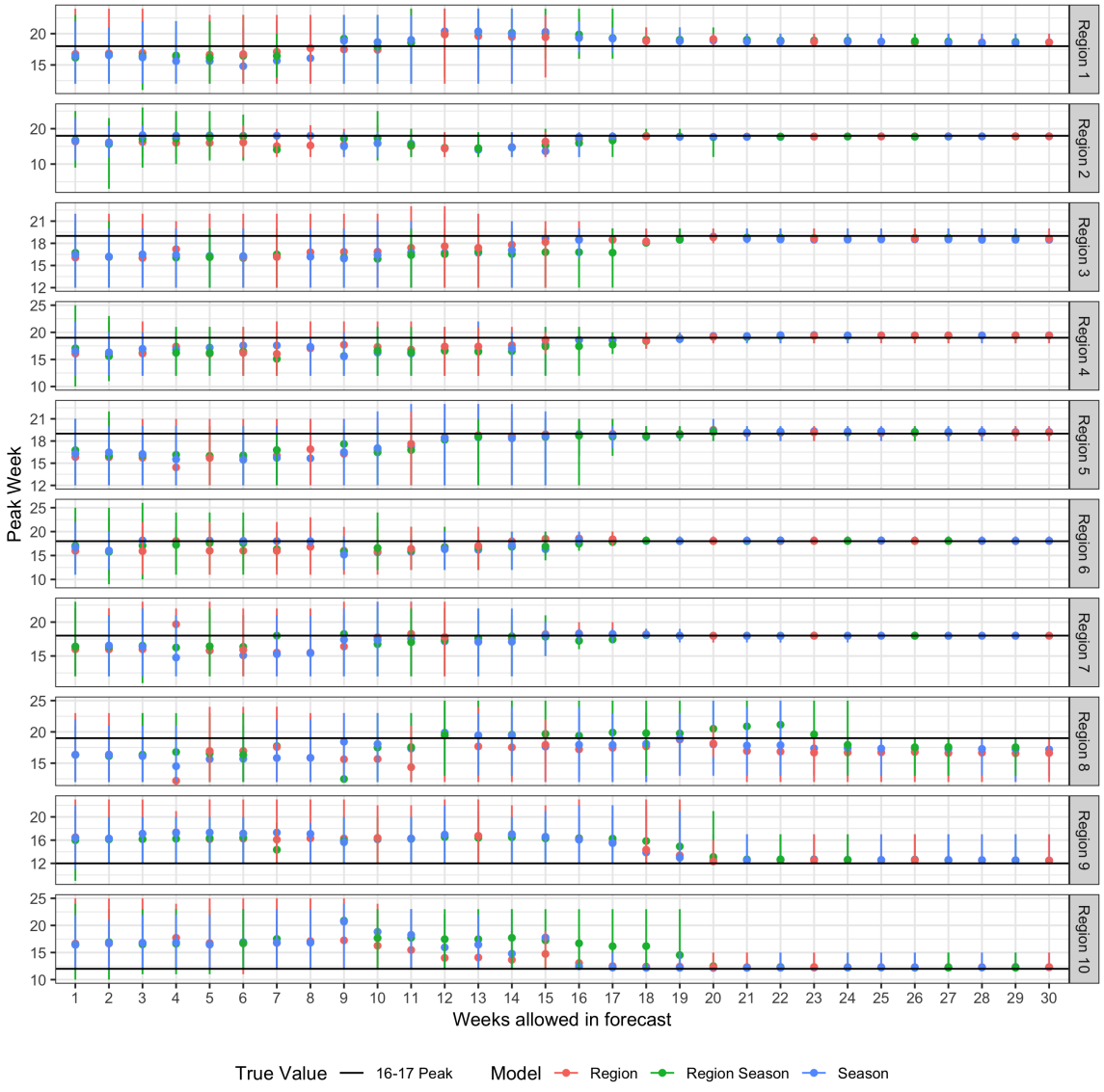

Figure 3.12 Estimated peak week of the forecasted season by how many weeks were given in the model. The colors represent the different hierarchy structures.

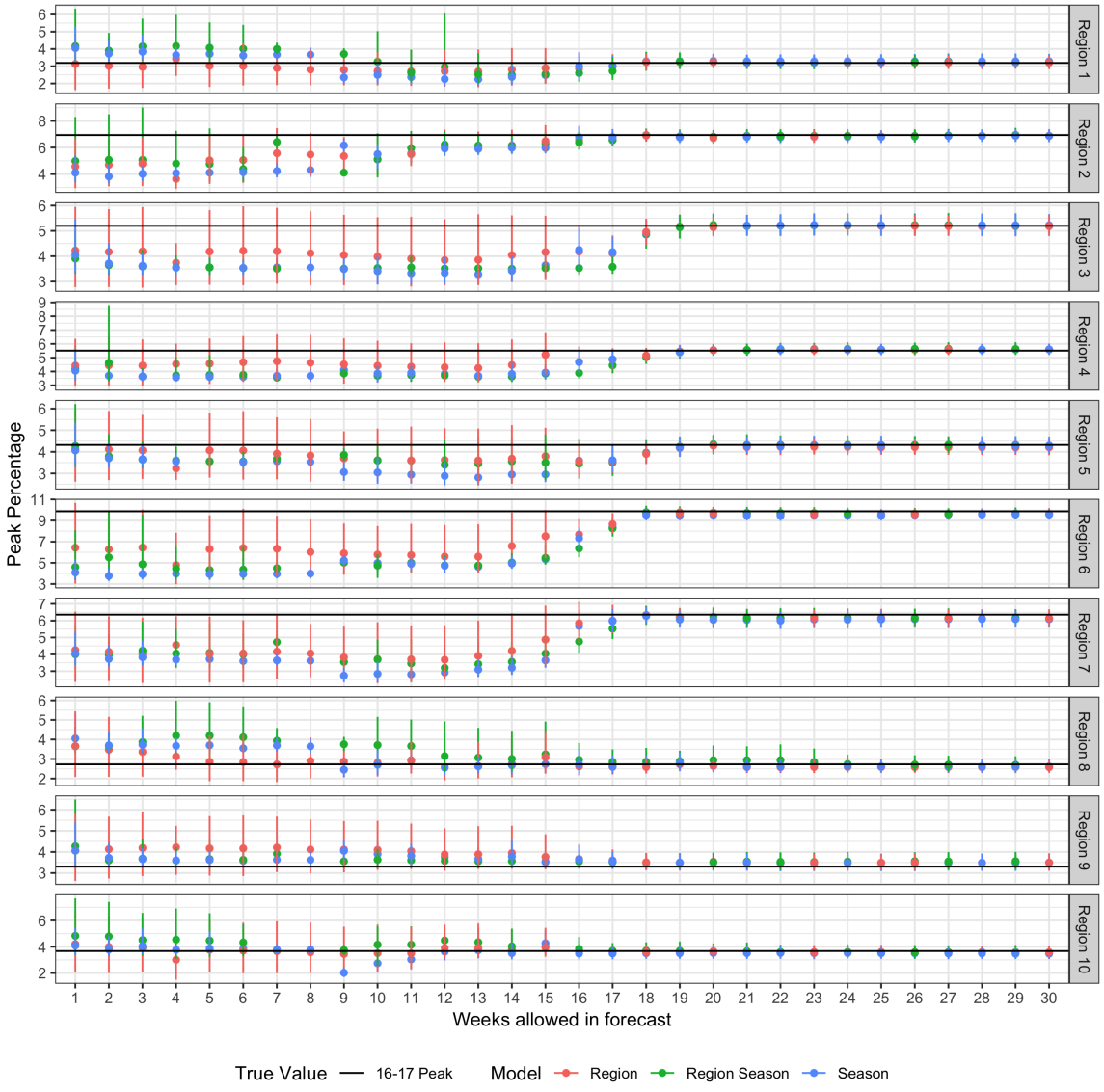

Figure 3.13 Estimated peak percentage of the forecasted season by how many weeks were given in the model. The colors represent the different hierarchy structures.

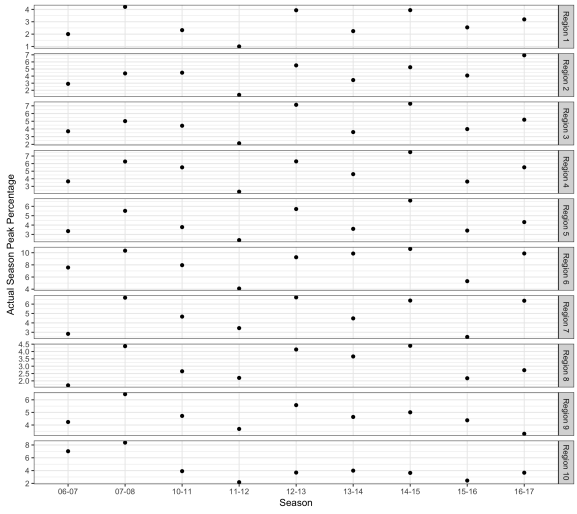

Figure 3.14 Observed peak percentages for all past seasons and all regions included in this analysis. Note that the scales are different for each row.

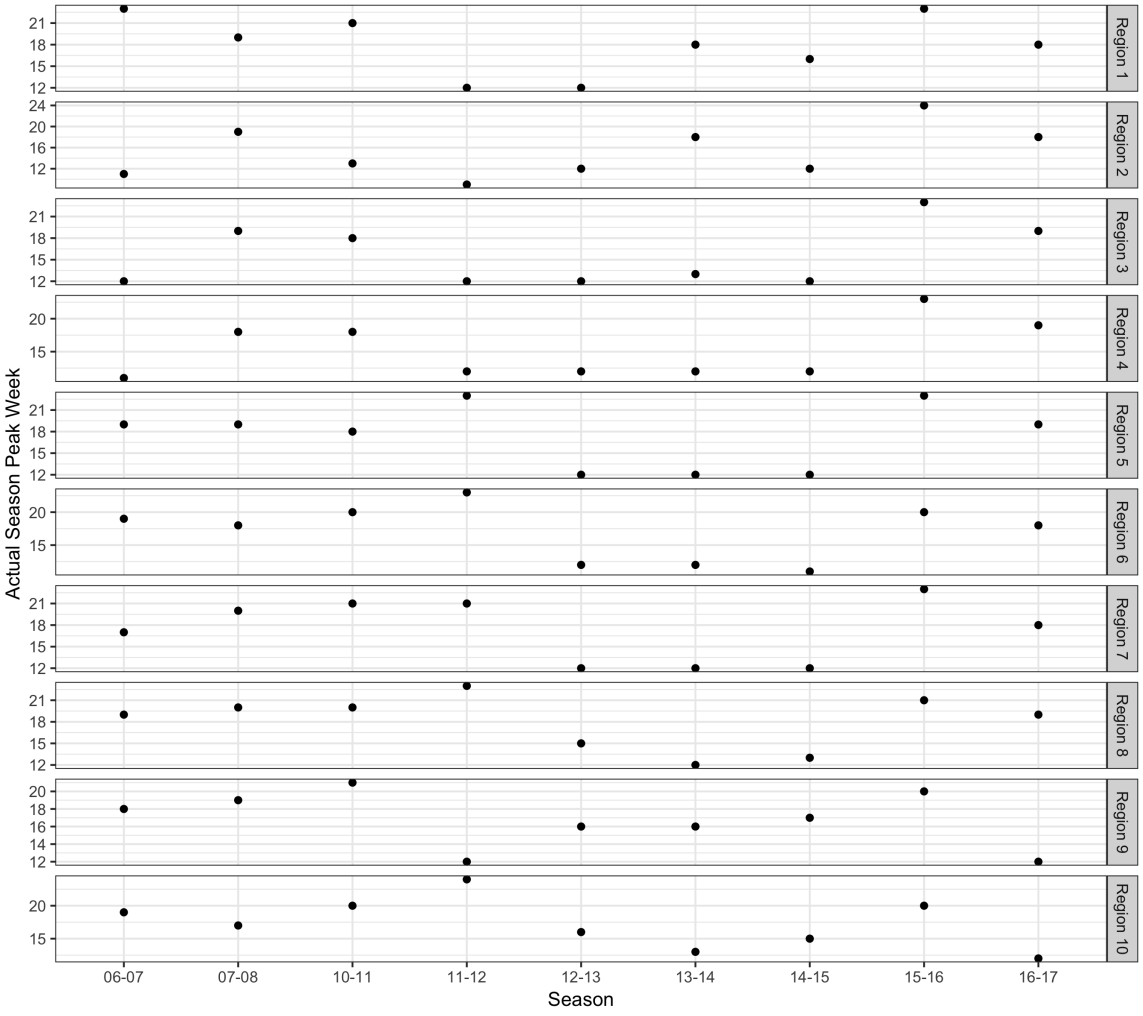

Figure 3.15 Observed peak weeks for all past seasons and all regions included in this analysis. Note that the scales are different for each row.

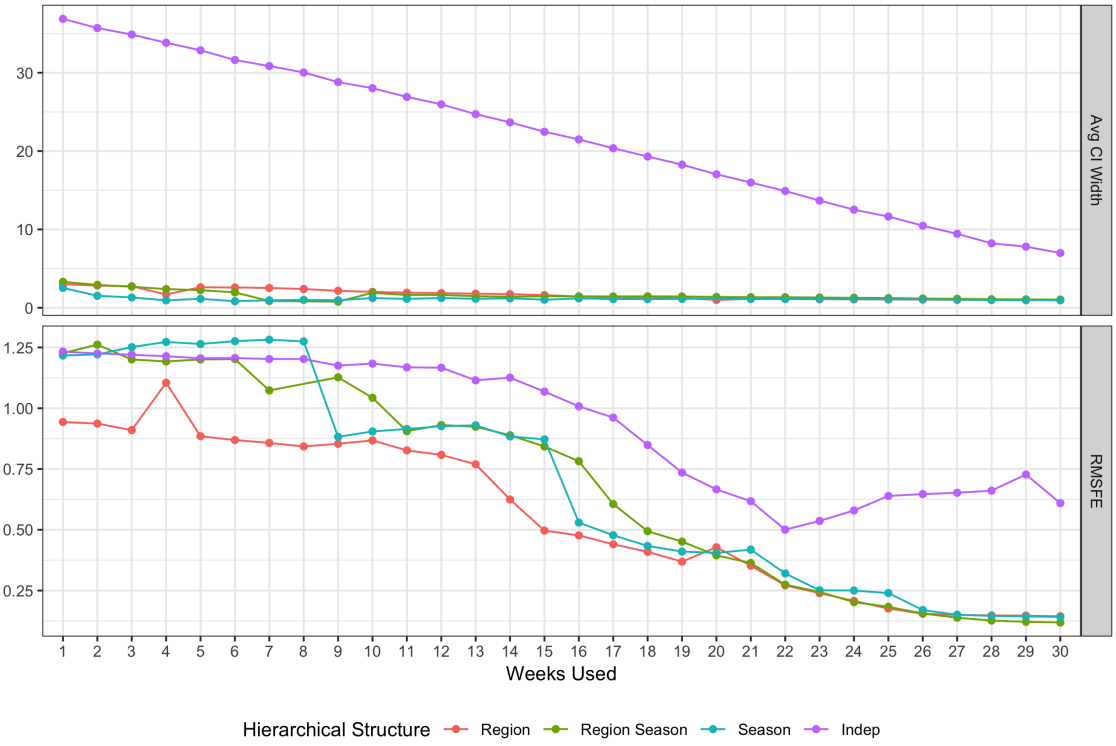

Figure 3.16 RMSFE and average 95% credible interval width for each hierarchical structure while using the hierarchical shrinkage prior by the number of weeks used in the forecast.

CHAPTER 4. DATA FUSION AND THE BENEFIT TO FORECASTING

4.1 Abstract

With the accessibility of the internet, internet searches have become a new source of data. Search data can be very valuable in areas where traditional data is not readily available. One potential problem with internet search data is the bias in the data. In this chapter, we look at supplementing influenza data from the CDC with Google search data. We conduct a simulation study to see if adding a second biased data source can help us forecast influenza. We compare two methods of modeling the mean: asymmetrical Gaussian functional form and functional principal analysis. Both methods use a hierarchical structure and are fit in a Bayesian framework. We find both methods preform comparably when recovering the true underlying mean. Adding a second data source is most beneficial when forecasting near the peak of the influenza season.

4.2 Introduction

Due to the accessibility of the internet, internet searches have become a potentially beneficial data source. Search data can be very valuable in areas where data is not easily accessible. For example, internet search data has been used to predict automotive sales and travel destinations (Choi and Varian, 2012). In both of these cases, datasets may be hard to come by but internet search data was able to supplement the lack of data and provide a helpful insights. Thanks to Google sharing its search query data through an API (trends.google.com, 2012) and `R` packages, such as `gtrendsR` (Massicotte and Eddelbuettel, 2018), interfacing with Google's API, internet search data is readily available. If traditional data sources exist, the problem lies in how to combine internet search data with more

traditional data sources in a coherent manner. One way to address this problem is with data fusion. Data fusion is the process of combining multiple sources of data in one coherent model for use in discovering the underlying data mechanism.

There are many different ways to implement data fusion and how the datasets are merged is dependent on the application. Some areas where data fusion is being implemented is genomics (Lanckriet et al., 2004), multisensor problems (Hall and Llinas, 1997), and remote sensing problems (Nguyen et al., 2012). Using internet searches in data fusion is also very prominent in studying and forecasting influenza (Stroup et al., 1988; Anderson, 2001; Soebiyanto et al., 2010; Corley et al., 2010; Culotta, 2010; Cook et al., 2011; Dugas et al., 2013; Paul et al., 2014; Xu et al., 2017; Michaud, 2016). Most of these approaches use either internet searches from Twitter (Culotta, 2010; Paul et al., 2014) or Google Flu Trends (Corley et al., 2010; Cook et al., 2011; Dugas et al., 2013; Michaud, 2016; Xu et al., 2017); others use climate data, or data from other diseases. Using Google Flu Trends has yielded great results in aiding forecasting, but Google shut down Google Flu Trends in 2015 (Google Flu Trends, 2015). In our approach, we will look at combining Google search data with ILINet data through data fusion to help us get a better nowcast (current forecast). Others have approached this problem but failed to include uncertainty quantification in their solution (Dugas et al., 2013; Paul et al., 2014; Michaud, 2016). Data fusion will be preformed using a multivariate model in which the mean of the model will either be a functional form or a linear combination of functional principal component. Both these approaches will be applied in a Bayesian hierarchical model which will provide uncertainty quantification through posteriors. The framework presented will be able to handle multiple data sources: Twitter mentions, Wikipedia searches, local hospital records, etc.

4.3 Data

For this project, we are interested in whether including multiple data sources can help us better learn and forecast the underlying true influenza rate curve. As mentioned earlier,

the literature shows that adding another data source to the ILINet data is beneficial in forecasting in the near future due to the way that ILINet data is captured.

4.3.1 ILINet

We use data from the CDC to study influenza. The data they provide cover people showing influenza-like illness (ILI). The Centers for Disease Control and Prevention (CDC) has formally defined ILI as "fever (temperature of $100°F[37.8°C]$ or greater) and a cough and/or a sore throat without a known cause other than influenza" (CDC, 2017). To collect and manage this data, the CDC has created the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). This is a network of 2800 outpatient healthcare providers throughout the United States of America and territories reporting the weekly number of patients they see in their office showing ILI each week and the total number of patients seen that week regardless of the reason. The weekly data is aggregated into regions by summing all patients with ILI in a region and summing all patients seen within a region.

We focus on the time period known as the influenza season. The CDC defines the influenza season as the weeks spanning Morbidity and Mortality Weekly Report (MMWR) weeks 40-20 roughly from November through early May. Since these weeks are primarily when the influenza rate changes the most, they are the focus. For ease of plotting and interpretation, the weeks have relabeled. Week 40 is relabeled to week 1 and week 20 is relabeled either week 32 or 33 depending on whether there were 52 or 53 MMWR weeks in the year.

Though this data is seen as the gold standard for estimating the true percentage of the population with ILI, ILINet data is produced by providers on a volunteer basis. They are not obligated to submit their data on time so consequently, ILINet data can be biased if not all the data is reported on time. Another concern is the potential lag between when the data is released and what week it correlates to. Due to the intricate systems the data needs to go through, it is possible for data that was released in week 12 to be a more accurate

picture of what happened in week 10. By using data from the internet, we may be able to get a more real time estimate of influenza rate.

4.3.2 Google Search Data

A popular search data source is Google Trends which reports the index of volume of Google searches by location. Google Trends does not provide the raw counts of searches but rather provides a query index. To create a query index, Google takes the total search volume for a search term in the location divided by the total number of searches in the location for some point in time. These relative volume search numbers are normalized so the start at 0 when they begin collecting data for that location and term (Choi and Varian, 2012). By looking into more specialized areas, we can create region data similar to ILINet data. Through careful conversion of dates to MMRW dates, we can convert Google search data to be on the same season-week timescale as ILINet. Figure 4.1 shows the weekly Google search data for the keyword 'influenza' and the CDC data for all regions in seasons 14-15 and 15-16. The data comes from the `gtrendsR` package (Massicotte and Eddelbuettel, 2018); it was aggregated from the United States and it looks very similar to ILINet data except it is more noisy. One potential problem with Google Trends data is that you are at the mercy of Google in which areas they are interested in monitoring or sharing the data so not all the states in a region may be available for the same amount of time. Nevertheless this data is free and thought to be an additional method to monitor the spread of disease.

To understand why Google search data may be a good indicator of the current influenza rate consider this: someone in your inner circle becomes sick with ILI so you to look up the symptoms of influenza as a preventative measure. This search hit is then a good indicator of influenza activity in the area. Unfortunately, this is not the only reason someone would look up the symptoms of a disease. One example of bias in the search data is someone looking up symptoms based on a national news report or a search for influenza symptoms

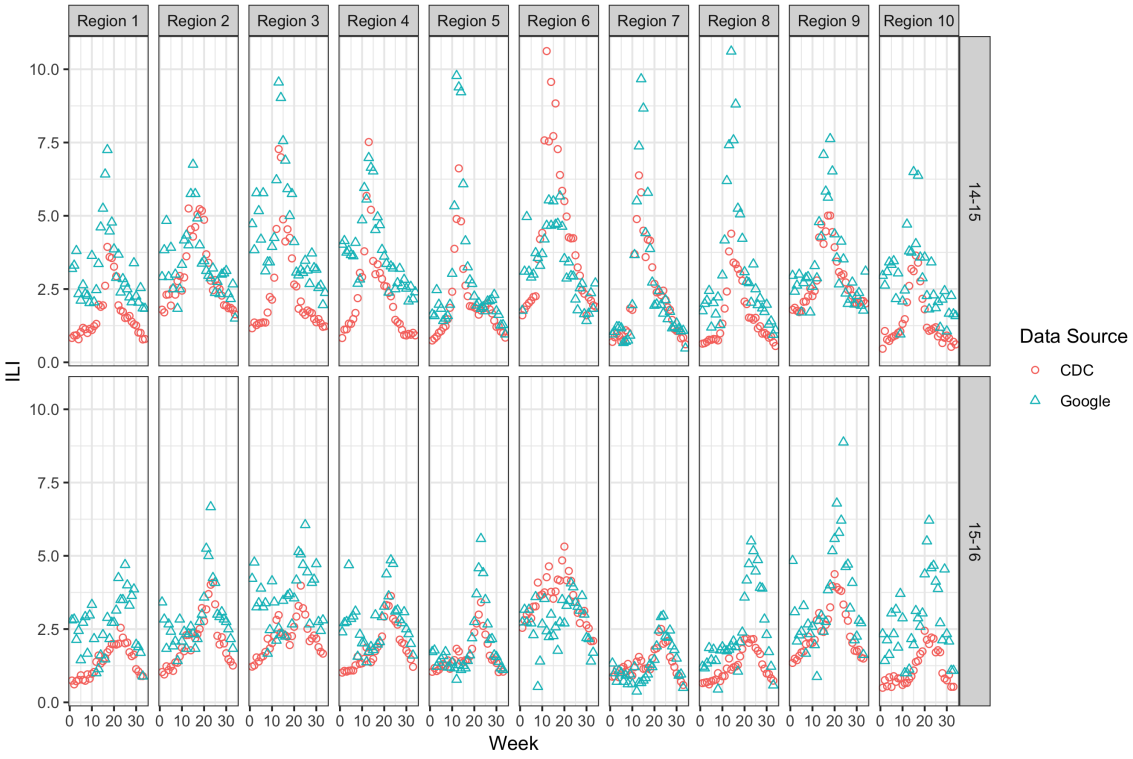

Figure 4.1 Weekly Google search data for the keyword 'influenza' and the CDC data for all regions in seasons 14-15 and 15-16. The data has been aggregated from searches across the United States. Data was gathered via the `gtrendsR` package.

based solely on the time of year. This is not a search directly linked to influenza activity in the area and therefore not a good indicator of influenza activity.

4.4 Methodology

In this section, we will describe the data model, asymmetrical Gaussian functional form, principal component analysis and hierarchical structures in these models. The methods used to assess model fit will be reviewed and the estimation details used in this analysis will be presented.

4.4.1 Data Model

A functional data model is presented in equation 4.1.

$$\begin{aligned} y_{r,s,1}(w) &= \psi_{r,s}(w) + \epsilon_{r,s,1}(w) \\ y_{r,s,2}(w) &= \psi_{r,s}(w) + \delta_{r,s,1} + \epsilon_{r,s,2}(w) \\ &\vdots \\ y_{r,s,d}(w) &= \psi_{r,s}(w) + \delta_{r,s,d-1} + \epsilon_{r,s,d}(w) \\ \epsilon_{r,s,i}(w) &\overset{ind}{\sim} N(0, \sigma^2_{\epsilon_i}) \end{aligned} \tag{4.1}$$

The observations, $y_{r,s,1}(w), \cdots, y_{r,s,d}(w)$, are compositions of the true underlying smooth function, $\psi_{r,s}(w)$, with some observation error, $\epsilon_{r,s,1}(w), \cdots, \epsilon_{r,s,d}(w)$, respectively and in all the sources but the first, i.e. $y_{r,s,2}(w), \cdots, y_{r,s,d}(w)$, a bias component, $\delta_{r,s,1}, \cdots, \delta_{r,s,d-1}$, which accounts for some of the difference in the responses from the underlying curve. We make the assumption with $y_{r,s,1}(w)$ that it is an unbiased source of data hence no need for a δ parameter. In this chapter, we will limit our studies to two data sources where one is treated as biased and one is not. We will compare two different methods to model the mean, $\psi_{r,s}(w)$: asymmetrical Gaussian functional form and Bayesian functional principal component.

4.4.2 Asymmetrical Gaussian Functional Form

More background of this approach can be found in the Chapter 2 of Ulloa (2019), but the basic intuition is that the region season mean, $\psi_{r,s}(w)$, will be modeled by a functional form created through two halves of normal densities joined at the mean of the two densities. The asymmetrical Gaussian functional form is based on the asymmetrical Gaussian (ASG) distribution which was first introduced by Fechner in 1897 (Wallis, 2014). We are not interested in the distribution, but rather a functional form of the distribution so the scaling

factor is swapped for something resembling the scaling factor in model 1 from Werker and Jaggard (1997). This results in the Asymmetrical Gaussian functional form presented in equation 4.2

$$ASG(w;\theta_{r,s}) = \begin{cases} \beta_1 + (\eta - \beta_1)\ \exp[-(w-\mu)^2/2\sigma_1^2] & w < \mu \\ \beta_2 + (\eta - \beta_2)\ \exp[-(w-\mu)^2/2\sigma_2^2] & w \geq \mu \end{cases} \tag{4.2}$$

$$\theta_{r,s} = (\beta_1, \beta_2, \eta, \mu, \sigma_1^2, \sigma_2^2)$$

For more modeling flexiblity, we model $\psi_{r,s}(w)$ on the logit scale. The ASG functional form hierarchical structure is presented in equation 4.3.

$$\begin{aligned} logit(\psi_{rs}(w)) &= ASG(w;\theta_{r,s}) \\ \theta_{r,s} &\overset{ind}{\sim} N(\theta_r, \Delta\Omega\Delta) \\ \theta_r &\overset{ind}{\sim} N(\theta, \Delta\Omega\Delta) \\ \Delta &= diag(\varsigma_1, \cdots) \end{aligned} \tag{4.3}$$

In order to model the $\theta_{r,s}$ jointly through a multivariate normal distribution, we took the log of the variance parameters, σ_1^2 and σ_2^2. The hierarchical structure presented above assumes the borrowing of information across seasons within a region to learn the region parameters, θ_r.

4.4.3 Bayesian Functional Principal Component Model

Functional principled component analysis (FPCA) was fully developed by Dauxois and Pousse (1976,1982) and for more information on this approach you may refer to Chapter 3 of Ulloa (2019). The FPCA method in equation 4.4 models the mean, $\psi_{r,s}(w)$ as a linear combination of principal components, v_k.

$$\psi_s(w) = \mu(w) + \sum_{k=1}^{\infty} \beta_{s,k}\mathrm{v}_k(w) \tag{4.4}$$

To choose the number of basis functions, v_k, the hierarchical shrinkage distribution will be used. Also proposed by Piironen and Vehtari (2017), the hierarchical shrinkage distribution is a simple twist on the horseshoe distribution. Instead of placing a half-Cauchy distribution on the λ parameters, a half-t with low degrees of freedom is used. Piironen and Vehtari (2017) claim the regularized horseshoe outperforms the hierarchical shrinkage distribution but in practice, we did not find this to be the case. The hierarchical shrinkage distribution is defined in equation 4.5.

$$\begin{aligned} \beta_{r,s,k}|\lambda_{r,s,k},\tau_{r,s} &\overset{ind}{\sim} N(0,\lambda^2_{r,s,k}\tau^2_{r,s}) \\ \lambda_{r,s,k} &\overset{ind}{\sim} t^+_v(0,1) \end{aligned} \tag{4.5}$$

In this model, $\lambda_{r,s,k}$ and $\tau_{r,s}$ represent the local and global shrinkage parameters, respectively. These parameters determine which explanatory variables are of consequence to the response.

Equation 4.6 combines the hierarchical shrinkage distribution with the data model.

$$\begin{aligned} y_{r,s}(w) - \widehat{\mu}(w) &= \sum_{k=1}^{K}\beta_{r,s,k}\mathrm{v}_k(w) + \epsilon_s(w) \\ \epsilon_s(w) &\overset{ind}{\sim} N(0,\sigma^2_\epsilon) \\ \beta_{r,s,k} &\overset{ind}{\sim} N(0,\lambda^2_{r,s,k}\tau^2_{r,s}) \\ \lambda_{r,s,k} &\overset{ind}{\sim} t^+_v(0,1) \end{aligned} \tag{4.6}$$

Just as the asymmetrical Gaussian functional form had a hierarchical structure on it, we can place one on the hierarchical shrinkage distribution. We can set up models that borrow information within each region and season i.e. each region-season combination is conditionally independent as in equation 4.7 ($\pi_1(\theta_0)$ and $\pi_2(\theta_0)$ are the respective distribution and prior on the shrinkage parameters). This gives each region-season combination the flexibility to choose its own principal components that are relevant to that particular

combination.

$$\tau_{r,s} \overset{ind}{\sim} \pi_1(\theta_0) \tag{4.7}$$
$$\lambda_{r,s,k} \overset{ind}{\sim} \pi_2(\theta_0)$$

Or we could have a similar shrinkage pattern within a region across the different seasons, replacing $\lambda_{r,s,k}$ and $\tau_{r,s}$ with $\lambda_{r,k}$ and τ_r, respectively and set up shrinkage parameters as in equation 4.8.

$$\tau_r \overset{ind}{\sim} \pi_1(\theta_0) \tag{4.8}$$
$$\lambda_{r,k} \overset{ind}{\sim} \pi_2(\theta_0)$$

Similarly, we the shrinkage pattern within a season across all regions could be the same by replacing $\lambda_{r,s,k}$ and $\tau_{r,s}$ with $\lambda_{s,k}$ and τ_s, respectively as in equation 4.9.

$$\tau_s \overset{ind}{\sim} \pi_1(\theta_0) \tag{4.9}$$
$$\lambda_{s,k} \overset{ind}{\sim} \pi_2(\theta_0)$$

$$\tag{4.10}$$

4.4.4 Simulated Data

Even though internet search data can be biased, we still want to examine how much of an advantage having a second observation can be. We have set up a simulation study with data that is similar to ILINet and Google search data. Figure 4.2 shows the two simulated data sources in different colors faceted by region and season. Both of the sources are deviations off of the underlying mean, the black line. The mean comes from an asymmetrical Gaussian functional form curve where the region curves follow a season mean curve since it looks like this is happening in the ILINet data. Just like ILINet and Google search data, we assume our two data sources are realizations of the same underlying process with more noise in the second source. Similar to ILINet data, we simulated ten different regions and since there is not much historical data available for internet searches, we limited the number of seasons to

five. The sources were created such that the second source would have a standard deviation five times greater that the first source. They both followed a normal distribution with ASG functional form curve mean and a standard deviation of 0.01 for source 1 and 0.05 for source 2.

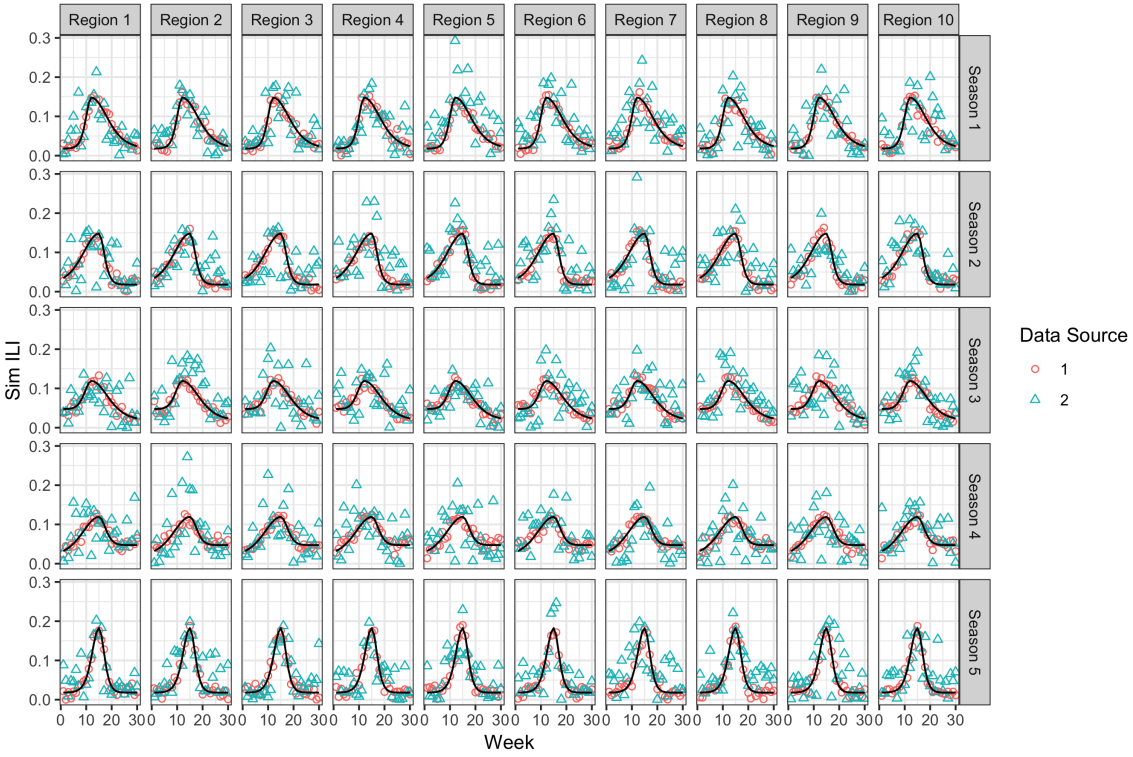

Figure 4.2 Two simulated sources in different colors are realizations of the same underlying mean curve (black line).

4.4.5 Registration

To create the principal components for the Bayesian FPCA model it is better if the data is centered. To do this, the weekly mean is calculated and each observation gets its respective mean subtracted from it. This process is outlined in equation 4.11 where i represent the data source i.e. $i \in (1, 2)$.

$$\mathrm{y}^{*}_{r,s,i}(w) = \mathrm{y}_{r,s,i}(w) - \widehat{\mu}(w,i) \tag{4.11}$$
$$\widehat{\mu}(w,i) = \frac{1}{RS}\sum_{r=1}^{R}\sum_{s=1}^{S}\mathrm{y}_{r,s,i}(w)$$

4.4.6 Model Checking

In this subsection, we list the remaining details needed to run these models and how we estimated the parameters in the models. The methods used to check the model fit are listed. Since we used Markov chain Monte Carlo (MCMC) to fit the models, we describe how to assess convergence of the MCMC. We then look at how to check the model fit and forecasting abilities.

4.4.6.1 Estimation

In order to estimate these models, we need prior distributions on some parameters. The priors for the Bayesian FPCA were minimal. Non-informative standard half-Cauchy priors were assigned to the data standard deviation (Gelman, 2006a; Polson and Scott, 2012) and the global shrinkage parameters. In the asymmetrical Gaussian model, we need to set these priors on ς, θ, and Ω. The priors are listed in equation 4.12.

$$\begin{aligned} \varsigma_i &\overset{ind}{\sim} t_4^{+}(0,1) \\ \theta &\overset{ind}{\sim} N(m_0, C_0) \\ \Omega &\overset{ind}{\sim} LKJ(1) \end{aligned} \tag{4.12}$$

The prior for θ is a combination of independent normal priors for the parameters in θ. Prior values of $m_0 = (0, 0, 2, 15, 2, 2)$ and C_0, a diagonal matrix of $(1, 1, 0.75, 2, 0.5, 0.5)$ provide a pretty vague starting curve. The standard deviations get similar standard priors in the form of independent standard half-t distributions with 4 degrees of freedom, and for the LKJ, prior we set ν equal to 1.

We also included an independent prior on the β parameters using no hierarchical structure or sparsity inducing distribution for comparison in equation 4.13. Each regression coefficient is assigned an independent vague normal prior.

$$\beta_{r,s,k} \overset{ind}{\sim} N(0, 10) \tag{4.13}$$

All models were fit using MCMC via `Stan` (Stan Development Team, 2016) through `R` (R Core Team, 2016). The ASG model ran for $10,000$ iterations with half used for burn-in, and the Bayesian FPCA model ran for $6,000$ iterations with half used for burn-in. Both models used one chain, though random starting points were used in the BFPCA model and maximum likelihood estimates (MLEs) were used as starting points for the ASG model.

4.4.6.2 Convergence Check

To assess for problems with convergence, we used Geweke's diagnostic. Geweke (1992) proposed a convergence diagnostic for Markov chains in which if there are no causes for concern, the Geweke statistics should look like draws from a standard normal distribution. Trace plots are also used in a visual inspection for issues with convergence.

4.4.6.3 Model Fit

Multiple models have been suggested so to understand which fit best fits the data and yields the most accurate forecasts we compare root mean squared error (RMSE) and root mean square forecast error (RMSFE). The formulas both are listed in equation 4.14. Since we are preforming a simulation study, instead of comparing to the observed data, we will compare our fit to the true underlying mean, $\psi_{r,s}(w)$. In the formulas, $\psi_s(w)$ is the true curve from region r, season s and week w; $\widehat{\psi_{r,s}}(w)$ is the estimated smooth underlying

function; and $\psi^*_{r,s}(w)$ and $\widehat{\psi^*_{r,s}}(w)$ are their counterparts in the forecasted season.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\psi_{r,s}(w) - \widehat{\psi_{r,s}}(w)\right)^2} \tag{4.14}$$

$$RMSFE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\psi^*_{r,s}(w) - \widehat{\psi^*_{r,s}}(w)\right)^2} \tag{4.15}$$

RMSE and RMSFE both present a measure of how close our predictions are the true underlying curve. We use the RMSE and RMSFE instead of the MSE and MSFE so they can be interpreted on the data scale.

4.5 Results

4.5.1 Convergence Check

Figure 4.3 plots the Geweke diagnostic qqplot for all parameters in the different models where the left plot shows the hierarchical shrinkage prior parameters and the right plot shows the ASG parameters. For the models used, the values in Figure 4.3 follow the straight line and there are no major causes for concern. This are consistent with results in earlier chapters.

4.5.2 Model Fit

Figure 4.4 shows the posterior mean fit and 95% credible intervals on the simulated data where each column represents a region and each row represents a season. The independent model and the region-season model have larger credible intervals than the rest but they all have similar mean fits. The independent model has such wide intervals that it is not included in Figure 4.4. Both simulated datasets are included and differ by shape; the true underlying curve is also plotted as a black line.

Figure 4.4 shows how well the FPCA and ASG functional form fit the data and how close they are the true underlying curve. The fitted models are almost indistinguishable from the true curve with the exception at the end of the season. There is more variability

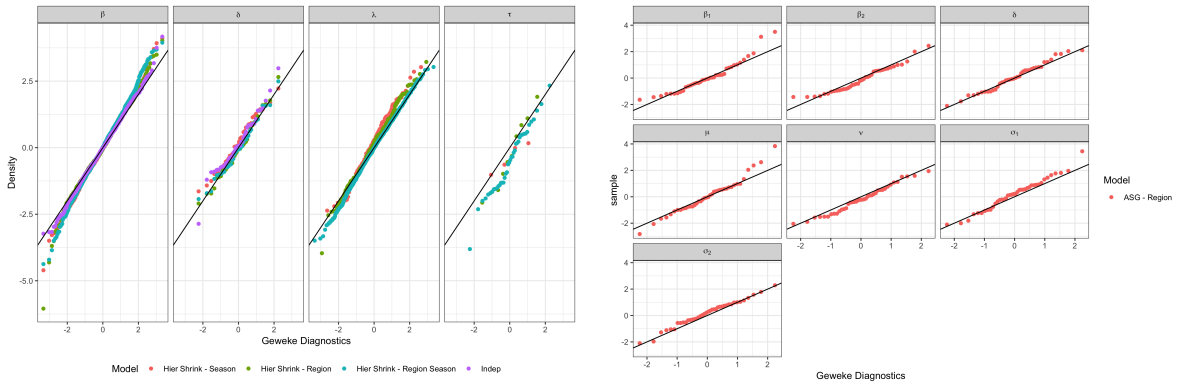

Figure 4.3 Geweke diagnostics Q-Q plot for the parameters of the different models. The left plot shows the hierarchical shrinkage prior parameters and the right plot shows the ASG parameters.

in the fitted curves. Table 4.1 shows just how good the fit is. The table shows the change in mean square error where the lowest mean square error is subtracted from the rest. The ASG model has the lowest RMSE, but all the models perform really well. Given the ASG functional form is the basis for the underlying curve it make sense that the ASG model fit the data the best. Likewise, the hierarchical shrinkage prior with the season hierarchical structure having the second lowest RMSE is not surprising given how the simulated data was created; the simulated data was created with season means so it should be that the season hierarchical structure is able to recognize that structure.

Table 4.1 The root mean square error of all the models across regions and seasons.

Model	Δ RMSE
ASG - Region	0
Hier Shrink - Season	0.001233674
Hier Shrink - Region	0.002165202
Hier Shrink - Region Season	0.003246454
Indep	0.041787257

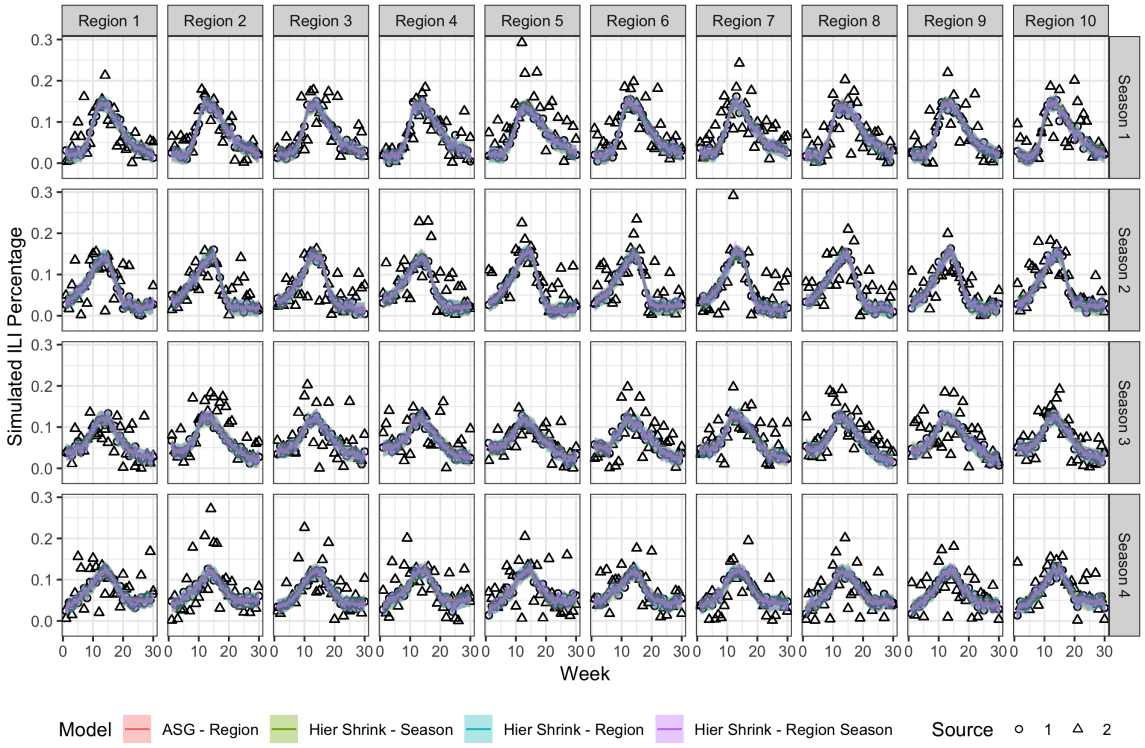

Figure 4.4 Posterior mean fit and 95% credible intervals on the simulated data for all models and regions and seasons.

4.5.3 Forecasting

What we are really interested in is not fitting the model to data, but rather forecasting and the effect of lag on our forecasts. We can look at our forecast by number of weeks allowed in the forecasts (Figures .10, 4.5, .11) or by lag (Figures 4.6, .12, .13, .14, .15).

For the figures focusing on the results by the weeks included in the forecast, (Figures .10, 4.5, .11) the rows are faceted by the number of lag weeks included and the columns are faceted by region. Both data sets are plotted in the figures and are denoted by the shapes where the triangles correspond to the noisier simulated data. As in Figure 4.2, the black line represents the true underlying curve.

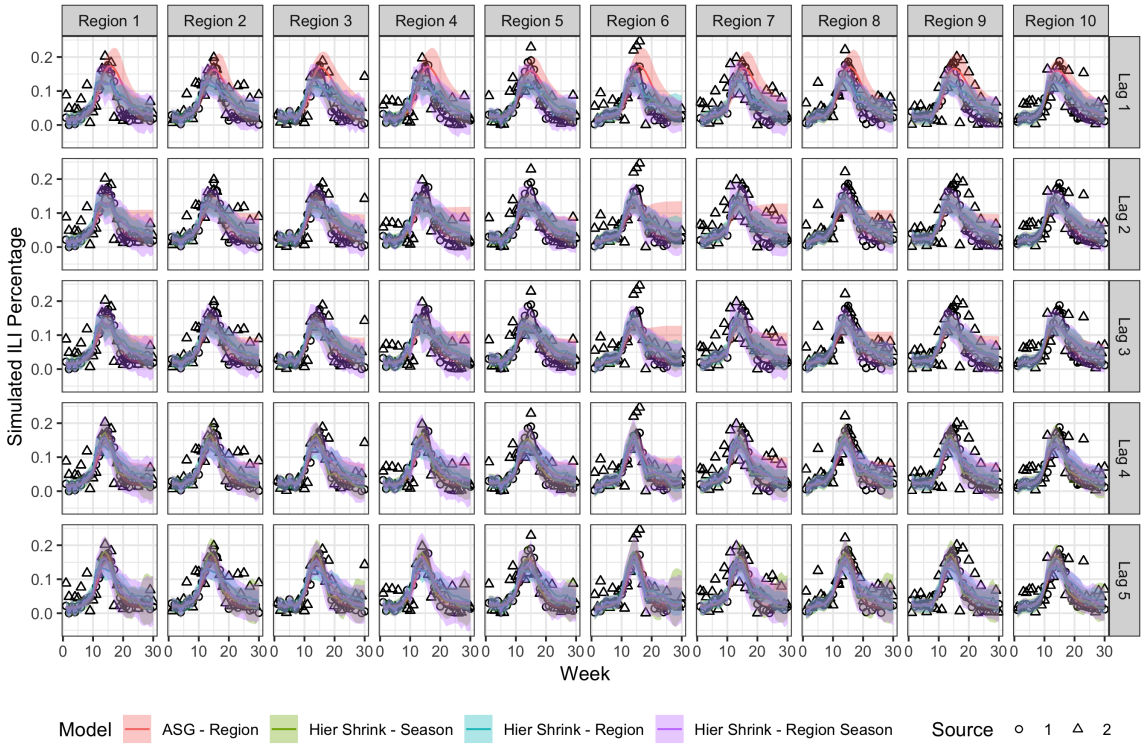

Figure 4.5 Posterior mean fit and 95% credible intervals on the simulated data when including 10 weeks of the forecasted season in the forecast. The columns are faceted by region and the rows are faceted by lag.

In Figures 4.5 and 4.6, it can be difficult to see how the lag affects the forecasts. In Figure 4.7, the log root mean square forecast error on the forecasted season is plotted by the number of weeks used as lag. The columns are faceted by the number of weeks used in the forecast. This plot allows us to see any minor changes that get brushed over in the large plots of the fits. The lag effect is not that great in either the five or fifteen week forecasts. The ten week forecast shows a big effect from the lagged weeks, especially with three to five lagged weeks. There is also a dramatic drop in RMSFE as the number of lagged weeks increases.

Figure 4.8, narrows in on the forecast error with forecasts using 10 weeks of data. The setup is the same as Figure 4.7 with the addition of forecast error from models using only

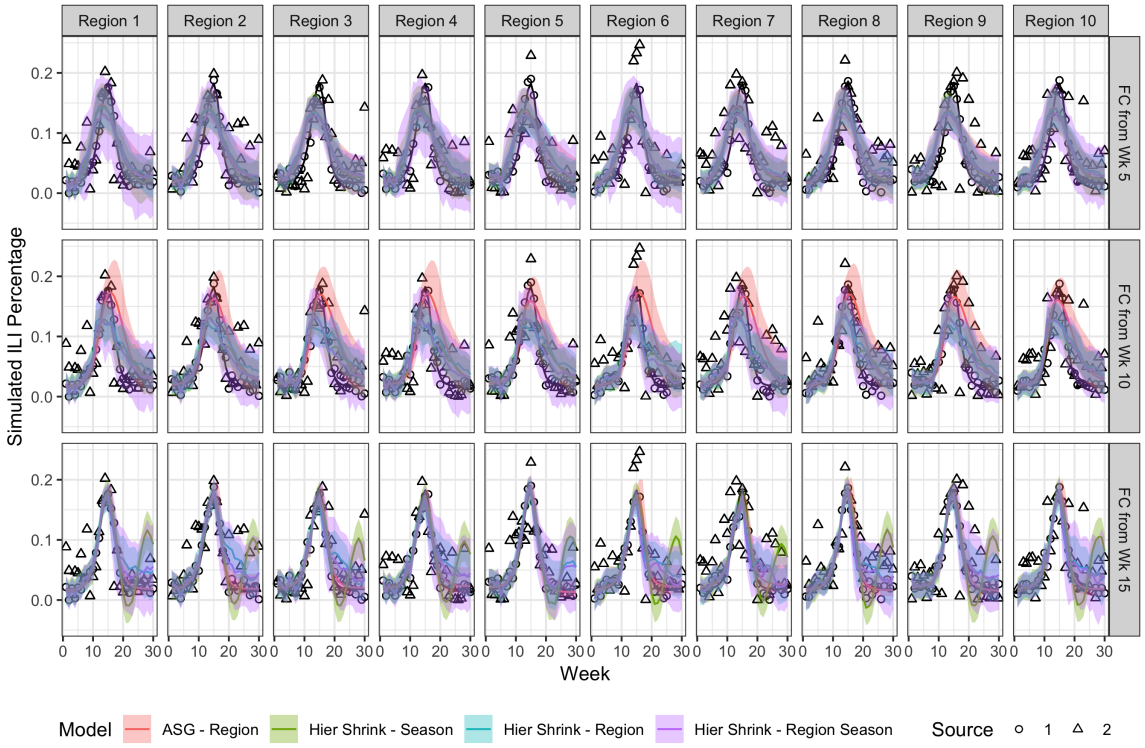

Figure 4.6 Posterior mean fit and 95% credible intervals on the simulated data when including 1 week of lag in the forecast. The columns are faceted by region and the rows are faceted by number of weeks included in the forecast.

one data source; those are the straight lines. This figure reiterates in usefuleness in adding a second source the closer you get to the peak. This indicates that short term forecasts around the peak would benefit the most from adding a second source.

Figures 4.10 and 4.9 show the estimated peak week and peak percentage by lag week while faceted by the number of weeks used in the forecast and regions on the rows. The colors denote the different models while the black line shows the true peak week and peak percentage on the respective plots. In both plots, you see the most change towards the true peak quantities when using ten weeks to forecast the new season and using three to five lag weeks. Just like with the RMSFE, the lag weeks are most beneficial the closer you get to the peak.

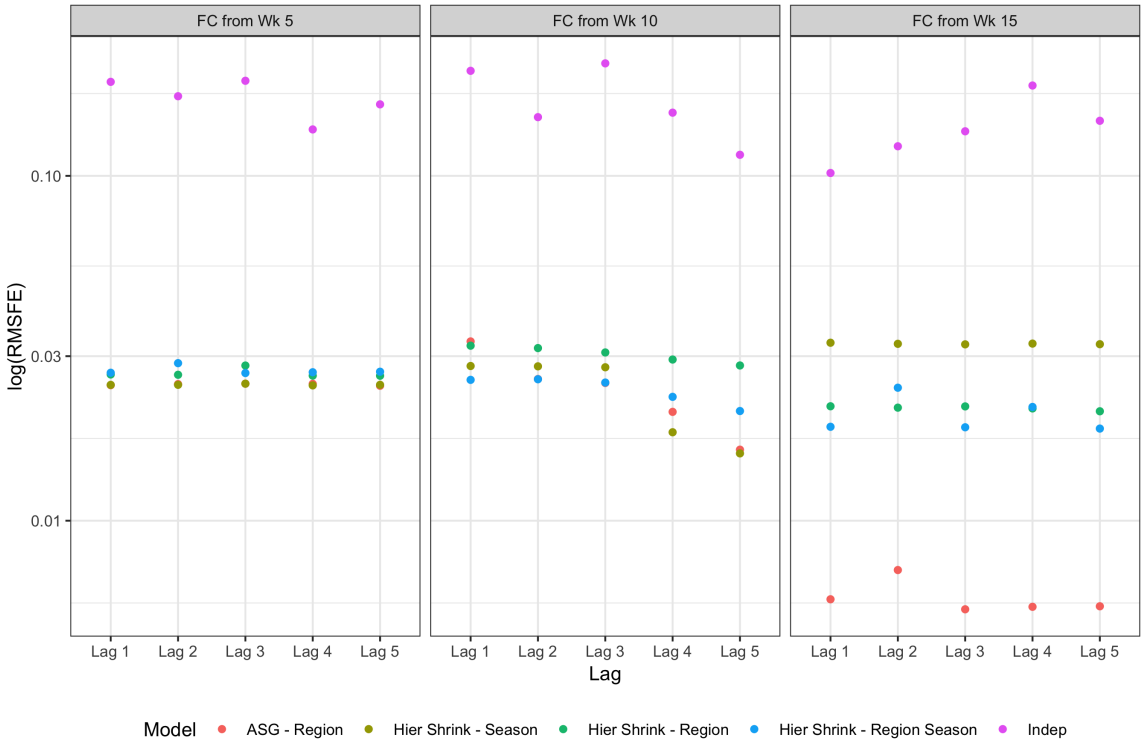

Figure 4.7 Log root mean square forecast error on the forecasted season. The columns are faceted by the number of weeks included in the forecast and the x axis is the number of lagged weeks included.

Focusing on the forecasted peak percentages when 10 weeks of data are included in the forecast, Figure 4.11 compares the peak percentage forecasts when using lag and using only one data source. The plot is similar to Firgure 4.9 with the addition of forecasts using only a single source of data plotted as a straight line. The forecasted peak percentage moves closer to the peak as you include more lag weeks improving on the forecasted peak week using only one source of data. The forecasts are using 10 weeks of data from source 1, so as more data is included from around the peak (lag 4 and 5) the forecasted peak percentages move closer to the truth. This highlights the benefit of the second source for short term week ahead forecasts around the peak.

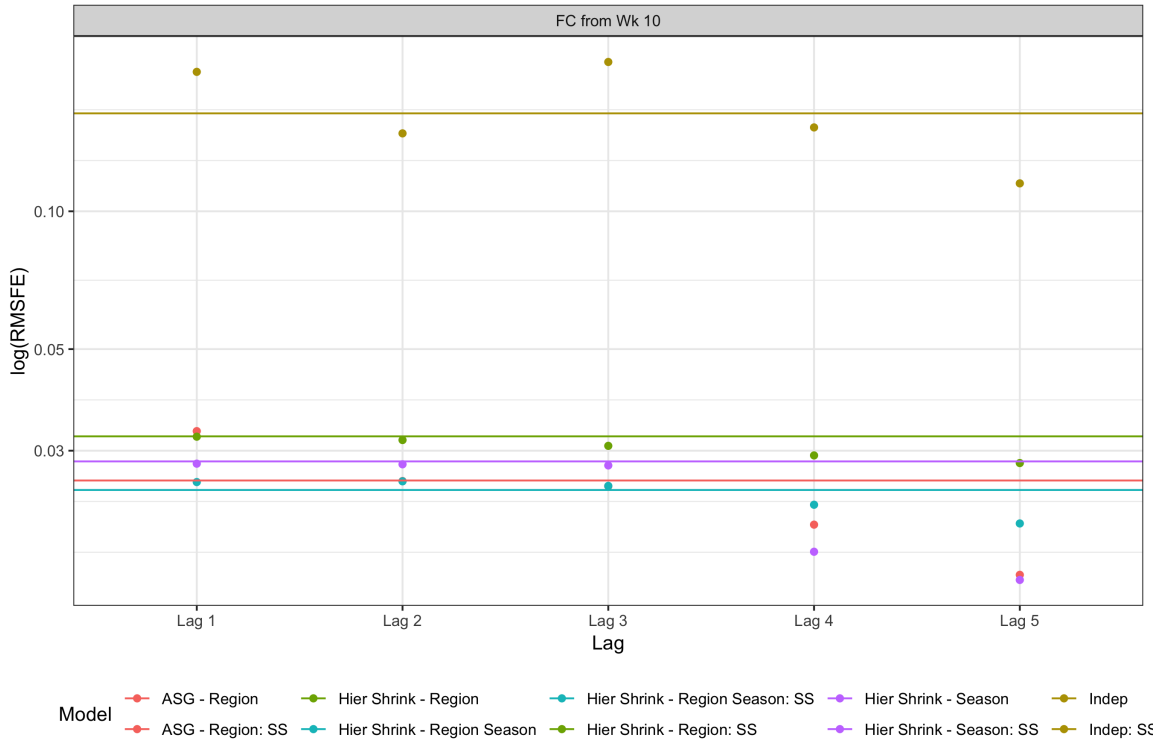

Figure 4.8 Log root mean square forecast error on the forecasted season with 10 weeks of data included in the forecast and the x axis is the number of lagged weeks included. The straight lines are also log root mean square forecast error except they only include one data source.

One assumption we made in the modeling stage was that the second source of data was going to be biased, but it would be accounted for by the $\delta_{r,s}$ random effect. Figure 4.12 shows the posterior means and 95% credible intervals for each $\delta_{r,s}$. The columns are faceted by region and the rows are faceted by the number of weeks included in the forecasted season. Focusing on season 5, we notice the estimates for $\delta_{r,s}$ change quite a bit as more weeks are included in the forecast: the intervals get much smaller and the mean estimates become more similar to the past seasons. Since there are bias parameters for each region season combination, it makes sense that as the forecasted season goes on, the estimates change quite dramatically. These δ parameters are picking up quite a bit of the bias; the mean

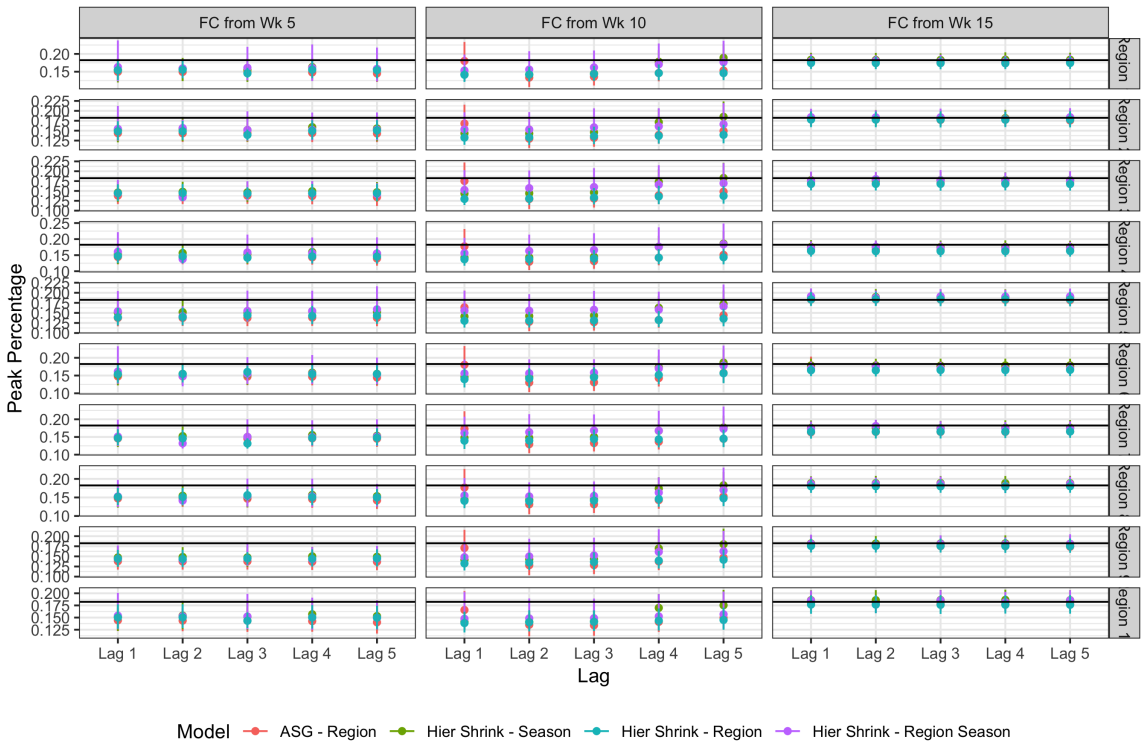

Figure 4.9 Estimated peak percentage by lag week and faceted by forecast week and regions. The models are denoted by color and the black line denotes the true peak percentage in $\xi_{r,s}(w)$.

estimates are hovering around 0.07 which is big considering the original scale of the data ranges from 0 to ≈ 0.2.

4.6 Conclusion

In this paper, we investigated how the lag between a reported measure and what that measure is currently affects the ability to forecast into the future. In our case, we are specifically interested in how internet search data can help us forecast influenza. We conducted a simulation study in which the datasets mimicked that of ILINet data and Google search data. In this simulation study, we examined the effect lag had on forecasting results. The models we used fit the data very well which was to be expected given how they performed

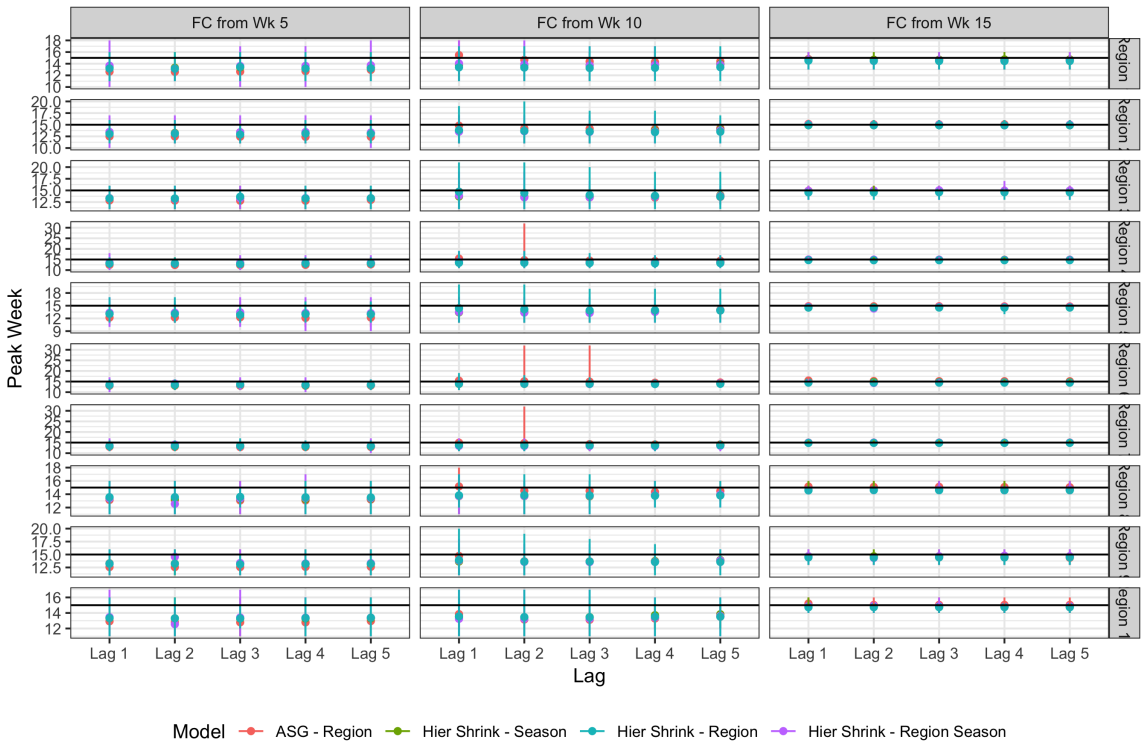

Figure 4.10 Estimated peak weak by lag week and faceted by forecast week and regions. The models are denoted by color and the black line denotes the true peak week in $\xi_{r,s}(w)$.

in earlier chapters. What was really of interest is how including a more recent data source to combat the lag in reporting affects the forecast. There was not a big effect when 5 or 15 weeks were included in the forecast, but when using 10 weeks in the forecast, the lag had an effect on improving accuracy. The effect was most evident with 3 to 5 weeks lag which corresponded to the peak of the season. The greatest benefit of having a second source was around when the peak of the influenza season would happen. When the peak occurs, we have little information on if it will it continue to grow or decrease. Therefore, any knowledge about what is happening around the peak is very beneficial. Around week 5, we are too early in the season and in week 15 we are past the peak of the season so as long as we know the rate is not continuing to increase, the forecast does not struggle much.

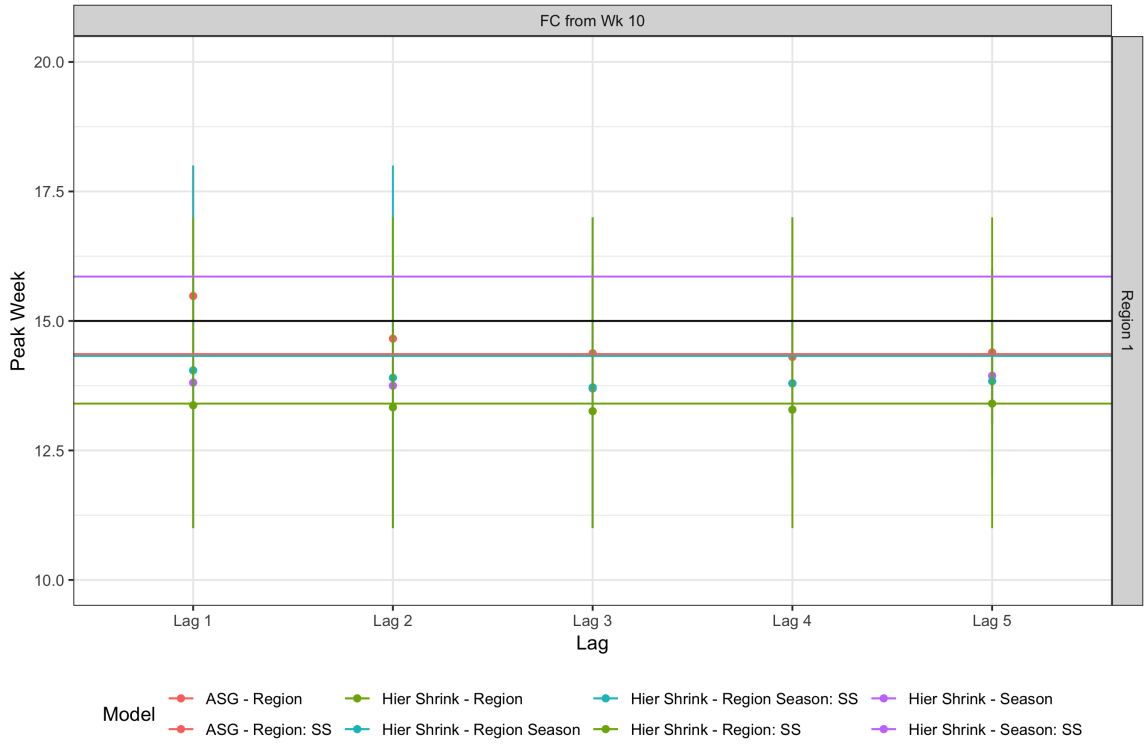

Figure 4.11 Estimated peak percentage by lag week using 10 weeks in the forecast for region 1. The models are denoted by color and the black line denotes the true peak week in $\xi_{r,s}(w)$. The straight lines are the estimated peak percentages for the forecasts using only one source.

Short term forecasts around the peak will see the biggest improvement in forecasting. This improvement came even with a biased second source of data. The modeling assumption of including a bias term, δ, was easily able to pick up on the bias of the second source. Long term forecasts need more work to take advantage of a second daa source as seen by the lack of improvement in RMSFE in forecasts using 5 weeks of data. These results are consistent with findings in Osthus et al. (citeyearOsthus19Internet).

Having a lag of four to five weeks is fairly dramatic, in real life we would expect the internet searches to be closer to catching up for one to two week lags. Future work would be to look at one to two week lags for more forecasts in between 10 and 15 weeks. This

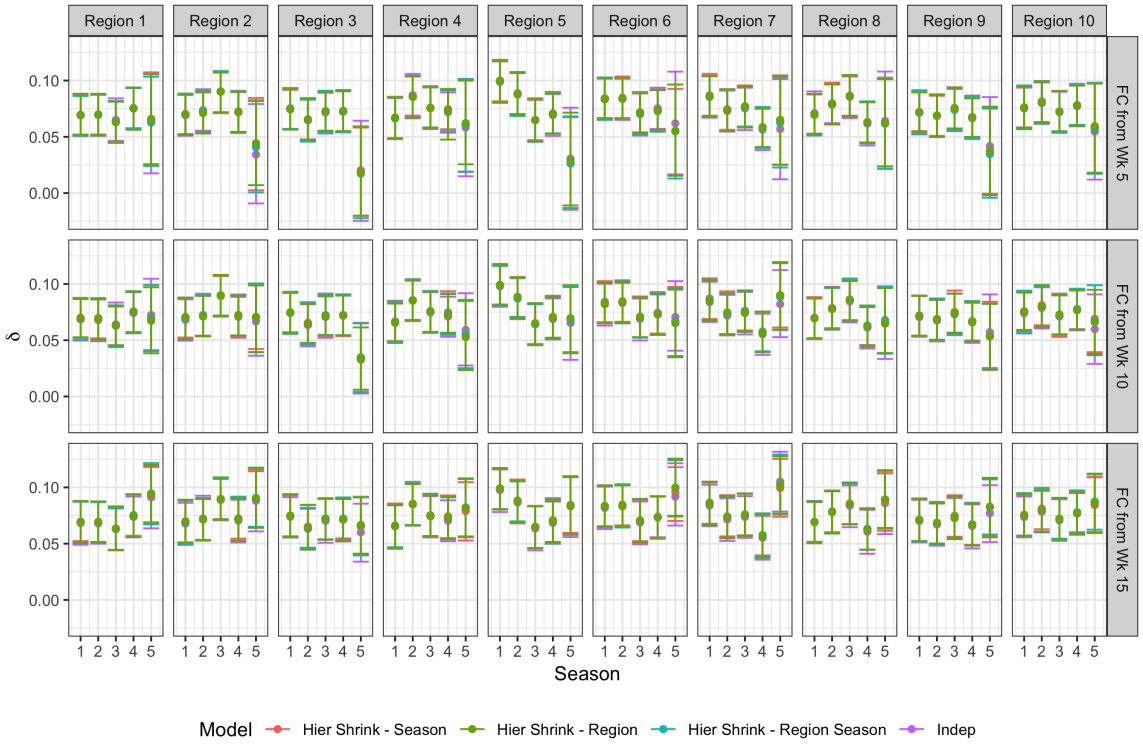

Figure 4.12 Posterior mean and 95% credible intervals plotted by season and faceted by region in the columns and number of weeks included in the forecast in the rows.

would really give us a sense of the practically of using internet data. We would also like to look at different bias terms. We included a region season bias term, but we could try including week, region, and season terms in the future. Lastly, we assumed one of our data sources was relatively unbiased, but we could just as easily assume it is biased and add a bias term to the model for all data sources.

CHAPTER 5. CONCLUSION

In this thesis, we explored a several methods to forecast the influenza season. Chapter 2 looked at using a functional form of the asymmetrical Gaussian distribution to model ILI. The asymmetrical Gaussian functional form was fit in a Bayesian hierarchical model where the hierarchical structures either borrowed information across regions or seasons. The asymmetrical Gaussian functional form provided a model with very interpretable parameters and flexibility. The necessity for the flexibility was studied to see if the functional form could be simplified to resemble a normal density, but by looking at posterior probabilities, we found the flexibility necessary to model influenza. This is because the time of the year we are interested in does not have time to reset to the baseline i.e. the starting and ending ILI percentage are not the same.

Chapter 3 switched the thinking of an observation being weekly percentages in a region-season to the set of percentages in a season being an observation i.e. a functional response. This allowed us to explore functional principal component analysis to model and forecast the influenza season. Functional principal component analysis considers each observation as the linear combination of principal components that are derived from the data. This model lacks interpretability in the parameters but provides flexibility and excellent forecasts. This method is able to reasonably forecast the peak week and peak percentage very early in the forecasted season. Through a fully Bayesian approach, shrinkage distributions are applied to parameter of each principal component. This allows us to navigate a common problem with the functional principal component methodology which is deciding how many principal components to use. By using the shrinkage distributions, the data decides which component coefficients to shrink towards 0 and which to allow to play a part. In our application, we found the shrinkage distributions limited us to around three to four components and shrink

the others. Multiple shrinkage distributions were compared. The horseshoe and regularized horseshoe distributions had convergence issues, but the hierarchical shrinkage distribution preformed well and showed no signs of concern with convergence. This chapter manages to tackle two big problems, long term forecasts and choosing number of principal components, with a fully Bayesian model based approach.

Chapter 4 uses the methods of chapters 2 and 3 and explores incorporating other data sources to assist in forecasting. Internet search data can give a better now-cast of influenza activity but can be biased. Our proposed model uses the methods in chapters 2 and 3 to model the mean of the data and employs a bias random effect for the second source of data. We preform a simulation study that mimics internet search data and ILINet data to examine the effect of lag and adding a second data source. We found that adding a second biased data source aided in forecasting especially around the peak of the season. Having a good idea of what is happening around the peak of the influenza season was very beneficial though this additional information was not as beneficial before or after the peak.

Chapters 2, 3 and 4 all employed hierarchical structures that either borrowed information across regions or season or sometimes both. When fitting the data, the seasonal effect was greater than regional effect. Though there was a regional effect, the responses were more similar across regions within a season than similar across regions within a season. However when forecasting, the regional hierarchical structures preformed better until there was enough data in the season for the season hierarchical structure to preform well. Lastly, there was not a dramatic improvement in the hierarchical models performance over an independent model when preforming inference. When looking at fitting the model to past data, both the independent structures and the hierarchical structures fit the data similarly though the hierarchical structures provide smaller credible intervals on the fit. Where the hierarchical structures really shine is in forecasting; the independent structures are more prone to poor mean fits and their credible intervals are much too wide when used to

forecast. The hierarchical structures provide much better mean fits and reasonable measures of uncertainty.

All chapters leave room for future work. In Chapter 2, more focus is needed on improving forecasts. Chapter 4 showed that this may be solved by adding a second data source. In Chapter 3, other basis functions that promote a smoother overall function should be considered. Chapter 4's future work entails studying in finer details when lag affects the most. We know around the peak is key. We would like to know exactly when. We also only used one data source, we would like to add more data sources and see if there is a significant gain in adding more than one additional source. With the other sources, it would be interesting to examine other bias terms, like week, region, and season bias terms. With all methods covered we want to improve our method of assessing the quality of uncertainty and computation time. Coverage in forecast credible intervals needs to be examined to see how reasonable the intervals are. All of these models took a considerable amount of time ($\approx$ 5 days). An empirical Bayes approach to these models might help to decrease computation time.

BIBLIOGRAPHY

Anderson, J. L. (2001). An ensemble adjustment kalman filter for data assimilation. *Monthly weather review*, 129(12):2884–2903.

Barnard, J., McCulloch, R., and Meng, X. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1311.

Betancourt, M. (2018). Bayes sparse regression.

Brockwell, P. J. and Davis, R. A. (2016). *Introduction to time series and forecasting*. Springer-Verlag Inc, Berlin; New York.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In van Dyk, D. and Welling, M., editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 73–80, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA. PMLR.

CDC (2011). Weekly u.s. influenza surveillance report.

CDC (2017). Weekly u.s. influenza surveillance report.

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Anal.*, 1(4):651–673.

Centers for Disease Control and Prevention (2019). Announcement of requirements and registration for the predict the influenza season challenge.

Choi, H. and Varian, H. (2012). Predicting the present with google trends. *Economic Record*, 88(s1):2–9.

Chowell, G., Sattenspiel, L., Bansal, S., and Viboud, C. (2016). Mathematical models to characterize early epidemic growth: a review. *Physics of life reviews*, 18:66–97.

Chretien, J.-P., George, D., Shaman, J., Chitale, R. A., and McKenzie, F. E. (2014). Influenza forecasting in human populations: a scoping review. *PloS one*, 9(4):e94130.

Cook, S., Conrad, C., Fowlkes, A. L., and Mohebbi, M. H. (2011). Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic. *PLOS ONE*, 6(8):1–8.

Corley, C. D., Cook, D. J., Mikler, A. R., and Singh, K. P. (2010). Using web and social media for influenza surveillance. In Arabnia, H. R., editor, *Advances in Computational Biology*, pages 559–564, New York, NY. Springer New York.

Crainiceanu, C. and Goldsmith, A. (2010). Bayesian functional data analysis using winbugs. *Journal of Statistical Software, Articles*, 32(11):1–33.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA. ACM.

Dauxois, J. and Pousse, A. (1976). *Les analyses factorielles en calcul des probabilit et en statistique: Essai d'tude synthtique*. PhD thesis, l'Universit Paul-Sabatier de Toulouse, France.

Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1):136–154.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *Ann. Appl. Stat.*, 3(1):458–488.

Dugas, A. F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T., and Rothman, R. E. (2013). Influenza forecasting with google flu trends. *PloS one*, 8(2):e56176.

Fechner, G. T. (1897). *Kollektivmasslehre*.

Gelman, A. (2006a). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1.

Gelman, A. (2006b). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Anal.*, 1(3):515–534.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *IN BAYESIAN STATISTICS*, pages 169–193. University Press.

Google Flu Trends (2015).

Greenspan, J. (2015). Preparing for ilinet 2.0. *Online Journal of Public Health Informatics*, 7(25).

Hall, D. L. and Llinas, J. (1997). An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6–23.

Hall, I., Gani, R., Hughes, H., and Leach, S. (2007). Real-time epidemic forecasting for pandemic influenza. *Epidemiology & Infection*, 135(3):372–385.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Hemmes, J. H., Winkler, K. C., and Kool, S. M. (1960). Virus survival as a seasonal factor in influenza and poliomyelitis. *Nature*, 188(4748):430–431.

Hickmann, K. S., Fairchild, G., Priedhorsky, R., Generous, N., Hyman, J. M., Deshpande, A., and Del Valle, S. Y. (2015). Forecasting the 20132014 influenza season using wikipedia. *PLOS Computational Biology*, 11(5):1–29.

Hyndman, R. J. and Shang, H. L. (2009). Forecasting functional time series. *Journal of the Korean Statistical Society*, 38(3):199 – 211.

Inst of Medicine (2006). The future of emergency care in the united states health system. *Annals of Emergency Medicine*, 48:115–120.

Lanckriet, G. R. G., Jordan, M. I., Cristianini, N., De Bie, T., and Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989 – 2001.

Linde, A. (2005). Dic in variable selection. *Statistica Neerlandica*, 59(1):45–56.

Longini Jr, I. M., Fine, P. E., and Thacker, S. B. (1986). Predicting the global spread of new infectious agents. *American journal of epidemiology*, 123(3):383–391.

M McDonnell, W., S Nelson, D., and E Schunk, J. (2011). Should we fear "flu fear" itself? effects of h1n1 influenza fear on ed use. 30:275–82.

MacDonald, I. L. and Zucchini, W. (1997). Hidden markov and other models for discrete-valued time series. London: Chapman & Hall.

Massicotte, P. and Eddelbuettel, D. (2018). *gtrendsR: Perform and Display Google Trends Queries*. R package version 1.4.2.

Michaud, N. L. (2016). *Bayesian models and inferential methods for forecasting disease outbreak severity*. PhD thesis, Iowa State University.

Mugglin, A. S., Cressie, N., and Gemmell, I. (2002). Hierarchical statistical modelling of influenza epidemic dynamics in space and time. *Statistics in Medicine*, 21(18):2703–2721.

Nguyen, H., Cressie, N., and Braverman, A. (2012). Spatial statistical data fusion for remote sensing applications. *Journal of the American Statistical Association*, 107(499):1004–1018.

Nsoesie, E., Mararthe, M., and Brownstein, J. (2013). Forecasting peaks of seasonal influenza epidemics. *PLoS currents*, 5.

Nsoesie, E. O., Brownstein, J. S., Ramakrishnan, N., and Marathe, M. V. (2014). A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and other respiratory viruses*, 8(3):309–316.

Osthus, D., Daughton, A. R., and Priedhorsky, R. (2019a). Even a good influenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited. *PLOS Computational Biology*, 15(2):1–19.

Osthus, D., Gattiker, J., Priedhorsky, R., and Del Valle, S. Y. (2019b). Dynamic bayesian influenza forecasting in the united states with hierarchical discrepancy (with discussion). *Bayesian Anal.*, 14(1):261–312.

Oviedo de la Fuente, M., Febrero-Bande, M., Muoz, M. P., and Domnguez, . (2018). Predicting seasonal influenza transmission using functional regression models with temporal dependence. *PLOS ONE*, 13(4):1–18.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Patwardhan, A. and Bilkovski, R. (2012). Comparison: Flu prescription sales data from a retail pharmacy in the us with google flu trends and us ilinet (cdc) data as flu activity indicator. *PLOS ONE*, 7(8):1–5.

Paul, M. J., Dredze, M., and Broniatowski, D. (2014). Twitter improves influenza forecasting. *PLoS currents*, 6.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Statist.*, 11(2):5018–5051.

Plummer, M. (2008). Penalized loss functions for bayesian model comparison. *Biostatistics*, 9(3):523–539.

Polson, N. G. and Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Anal.*, 7(4):887–902.

Priestley, M. B. (1978). Non-linear models in time series analysis.

Quenel, P. and Dab, W. (1998). Influenza a and b epidemic criteria based on time-series analysis of health services surveillance data. *European journal of epidemiology*, 14(3):275–285.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Shaman, J. and Karspeck, A. (2012). Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430.

Shaman, J. and Kohn, M. (2009). Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences*, 106(9):3243–3248.

Shang, H. L., Wisniowski, A., Bijak, J., Smith, P. W., and Raymer, J. (2013). Bayesian functional models for population forecasting. In *Joint Eurostat/UNECE Work Session on Demographic Projections*.

Soebiyanto, R. P., Adimi, F., and Kiang, R. K. (2010). Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. *PloS one*, 5(3):e9450.

Stan Development Team (2016). RStan: the R interface to Stan. R package version 2.14.1.

Stan Development Team (2018). Stan modeling language users guide and reference manual, version 2.18.0.

Stroup, D. F., Thacker, S. B., and Herndon, J. L. (1988). Application of multiple time series analysis to the estimation of pneumonia and influenza mortality by age 1962–1983. *Statistics in medicine*, 7(10):1045–1059.

Thompson, W., Shay, D., Weintraub, E., and et al (2004). Influenza-associated hospitalizations in the united states. *JAMA*, 292(11):1333–1340.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

trends.google.com (2012). Google trends.

Ulloa, N. (2019). *Bayesian hierarchical models for forecasting influenza*. PhD thesis, Iowa State University.

Vehtari, A. and Gelman, A. (2014). Waic and cross-validation in stan .

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432.

Viboud, C., Bjørnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., and Grenfell, B. T. (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*, 312(5772):447–451.

Wallis, K. F. (2014). The two-piece normal, binormal, or double gaussian distribution: Its origin and rediscoveries. *Statist. Sci.*, 29(1):106–112.

Wand, M. (2018). *SemiPar: Semiparametic Regression*. R package version 1.0-4.2.

Wang, J.-L., Chiou, J.-M., and Mueller, H.-G. (2015). Review of functional data analysis.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, 11:3571–3594.

Werker, A. R. and Jaggard, K. W. (1997). Modelling Asymmetrical Growth Curves that Rise and then Fall: Applications to Foliage Dynamics of Sugar Beet (Beta vulgaris L.) . *Annals of Botany*, 79(6):657–665.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models (2Nd Ed.)*. Springer-Verlag, Berlin, Heidelberg.

WHO (2018). Influenza (seasonal).

Xu, Q., Gel, Y. R., Ramirez, L. L. R., Nezafati, K., Zhang, Q., and Tsui, K.-L. (2017). Forecasting influenza in hong kong with google search queries and statistical model fusion. *PloS one*, 12(5):e0176690.

Yu, H., Alonso, W. J., Feng, L., Tan, Y., Shu, Y., Yang, W., and Viboud, C. (2013a). Characterization of regional influenza seasonality patterns in china and implications for vaccination strategies: Spatio-temporal modeling of surveillance data. *PLOS Medicine*, 10(11):1–16.

Yu, H., Alonso, W. J., Feng, L., Tan, Y., Shu, Y., Yang, W., and Viboud, C. (2013b). Characterization of regional influenza seasonality patterns in china and implications for vaccination strategies: Spatio-temporal modeling of surveillance data. *PLOS Medicine*, 10(11):1–16.

APPENDIX. ADDITIONAL MATERIAL

The appendix holds additional plots from chapters 2 and 4.

Plots from Chapter 2

Here are some additional plots referenced in the paper.

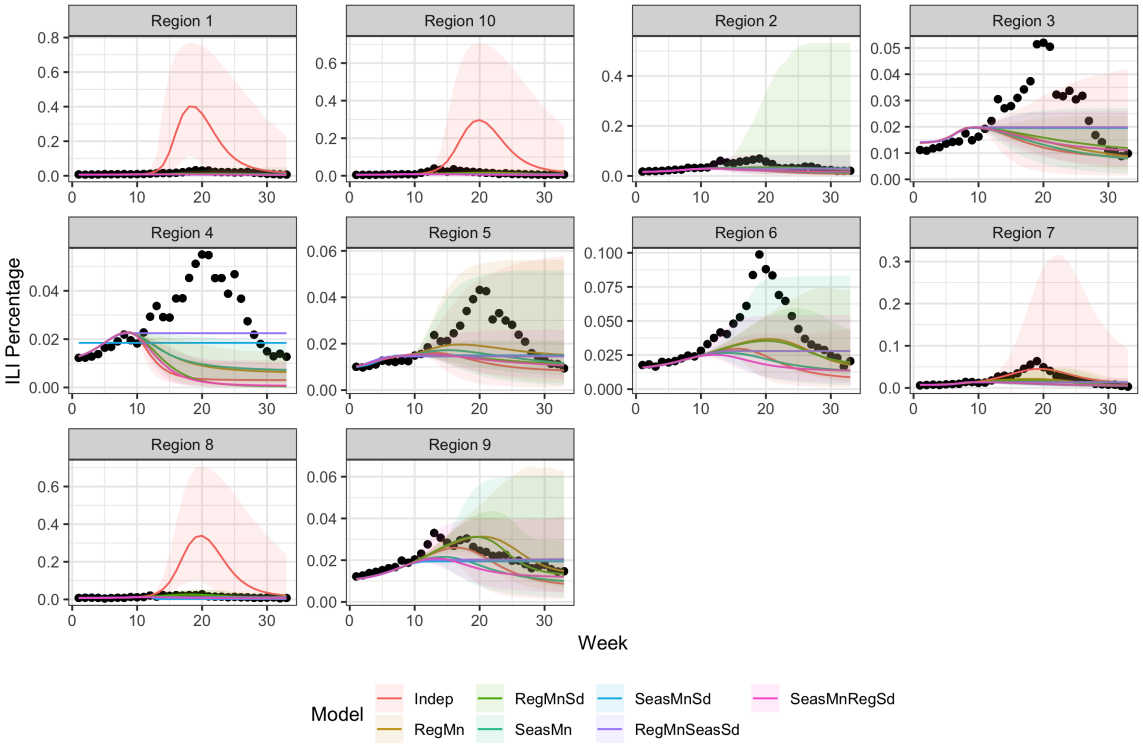

Figure .1 Long-term forecasts of all hierarchical structures for all regions in the 2016-2017 influenza season including only 10 weeks of data from the forecasted season.

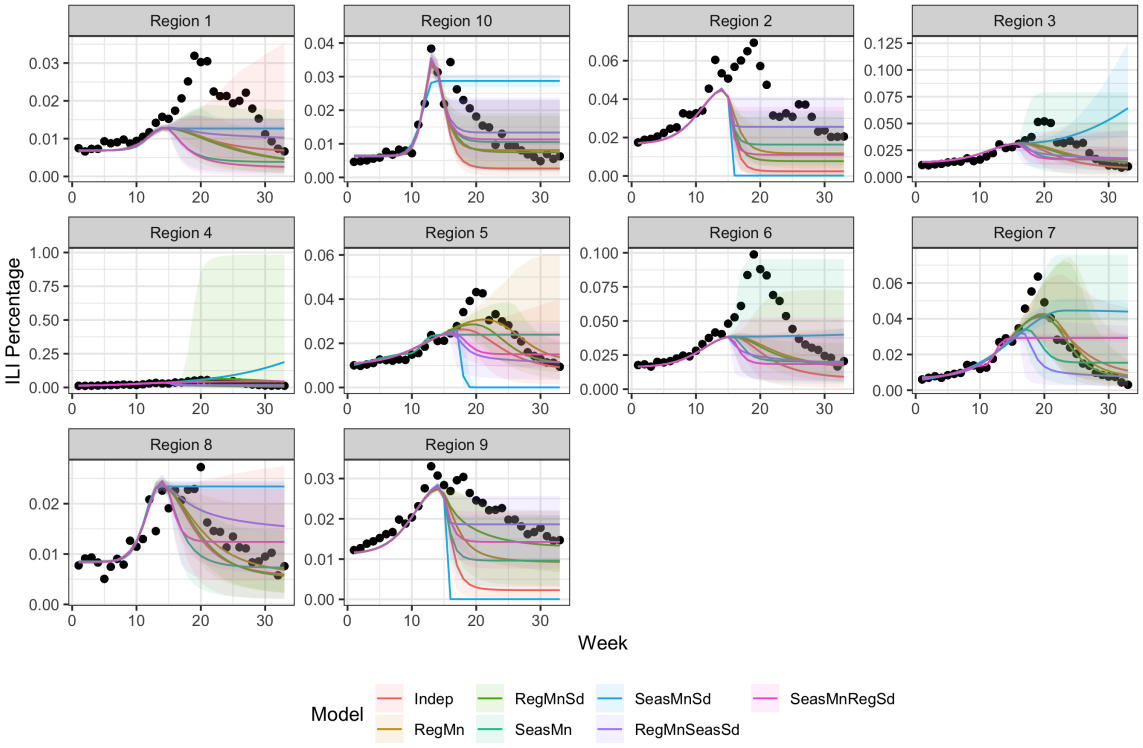

Figure .2 Long-term forecasts of all hierarchical structures for all regions in the 2016-2017 influenza season including only 15 weeks of data from the forecasted season.

Plots from Chapter 4

Here are some additional plots from our analysis.

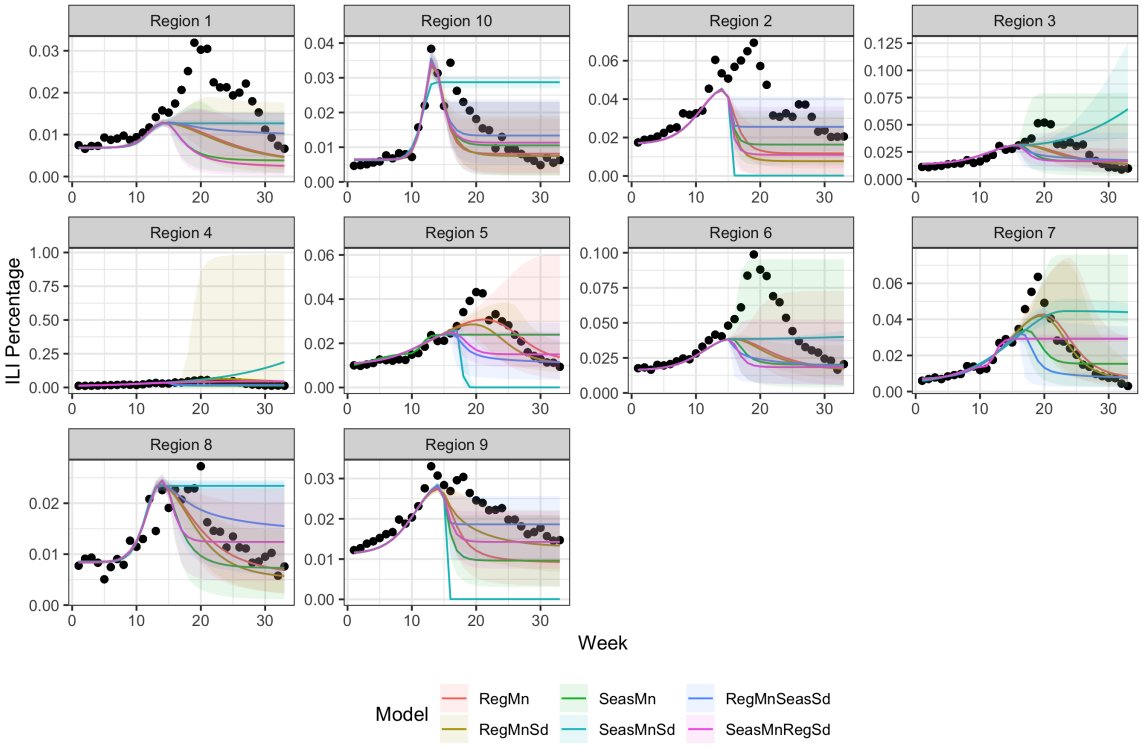

Figure .3 Long-term forecasts of all hierarchical structures excluding the independent model for all regions in the 2016-2017 influenza season including only 15 weeks of data from the forecasted season.

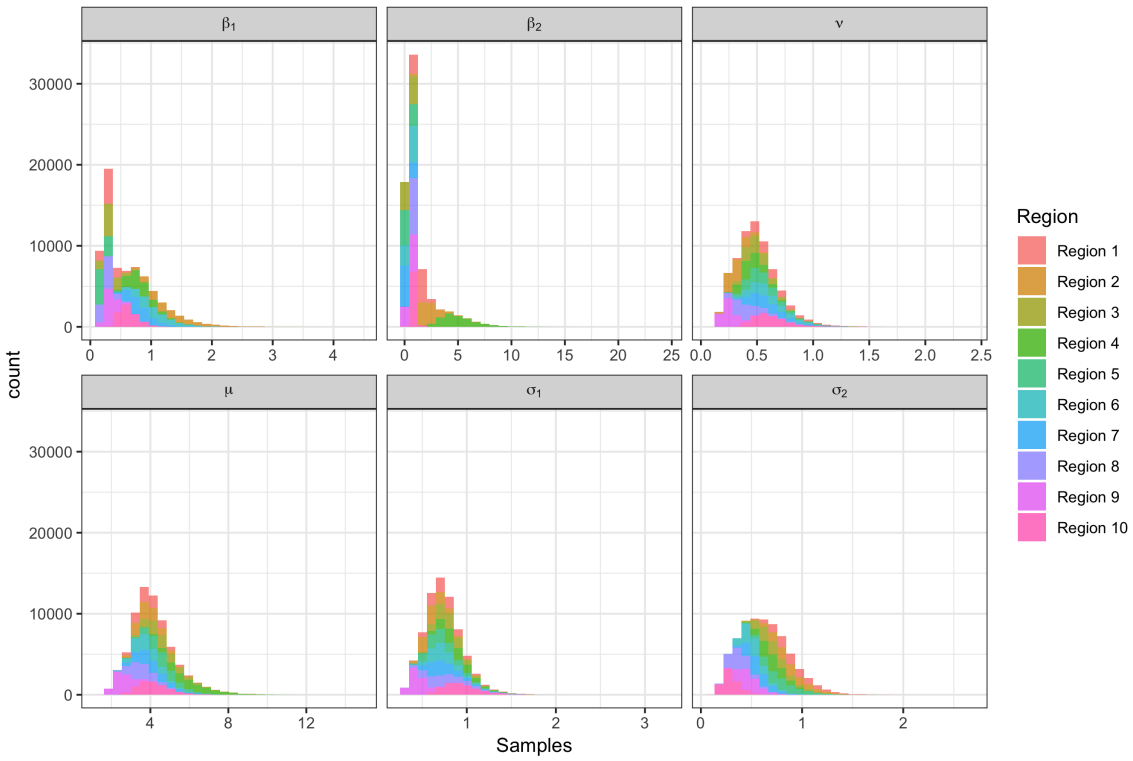

Figure .4 Posterior densities of the parameters in the Asymmetrical Gaussian functional form for the hierarchical structure using a region mean and standard deviation structure using ten weeks for forecasting.

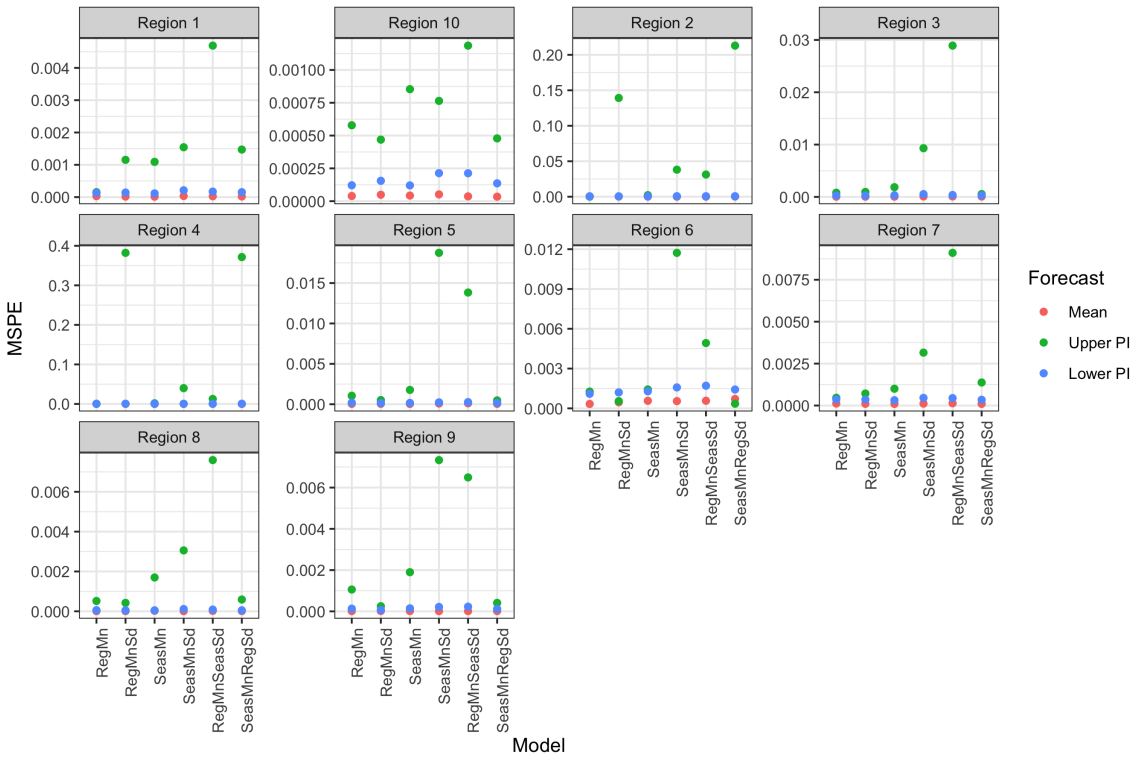

Figure .5 MSFE for the posterior mean forecast and their 95% credible intervals for all hierarchical structures using 3 weeks of data in the forecasted season.

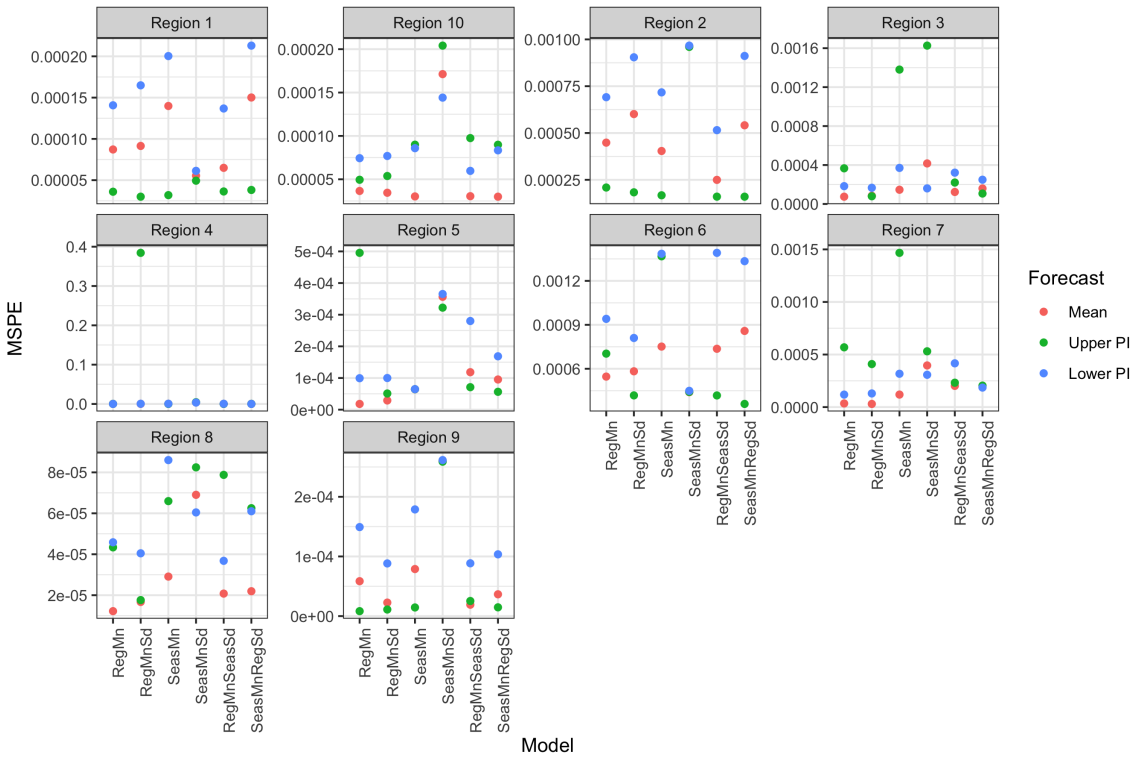

Figure .6 MSFE for the posterior mean forecast and their 95% credible intervals for all hierarchical structures using 15 weeks of data in the forecasted season.

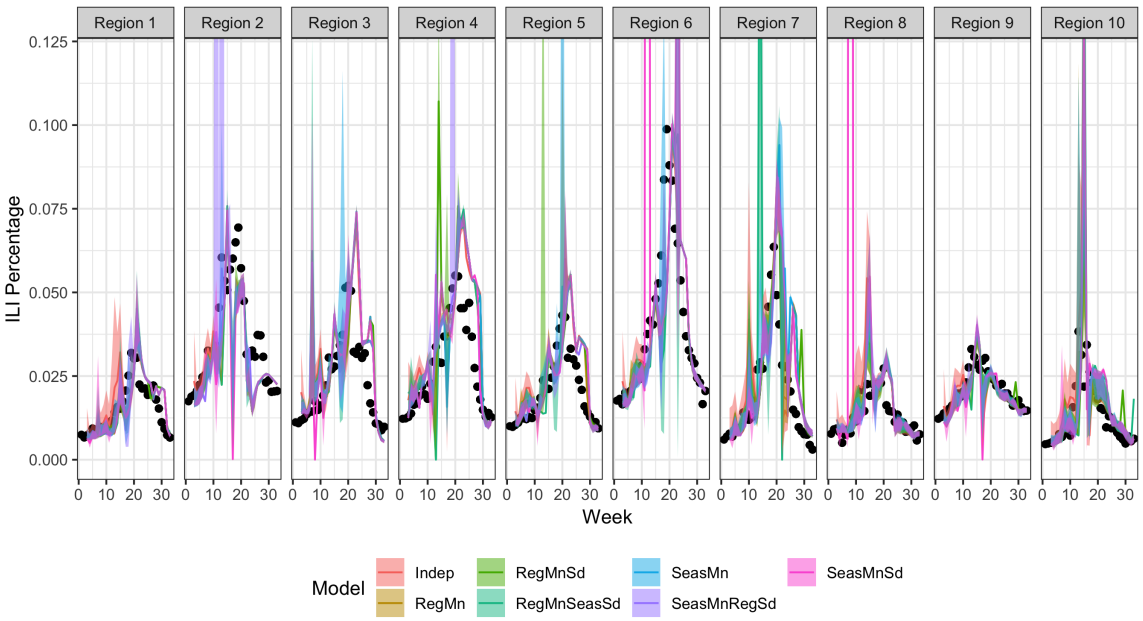

Figure .7 2 week ahead forecasts and their 95% credible intervals for all regions in the $2016 - 2017$ influenza season.

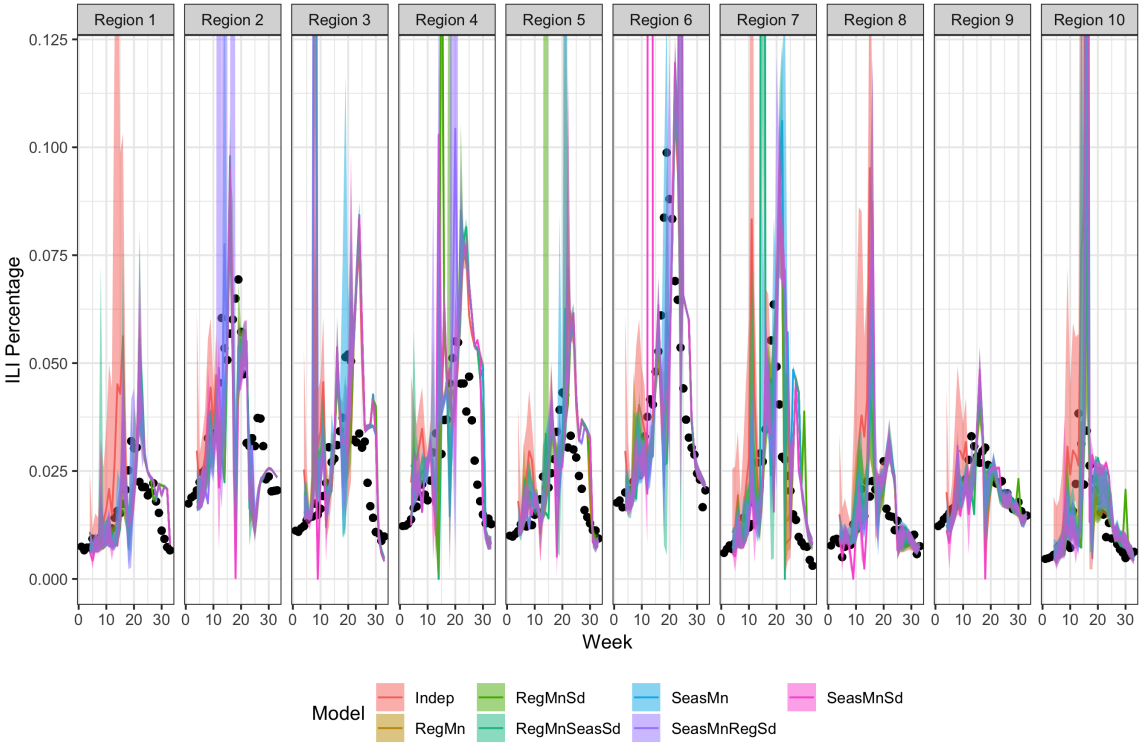

Figure .8 3 week ahead forecasts and their 95% credible intervals for all regions in the $2016 - 2017$ influenza season.

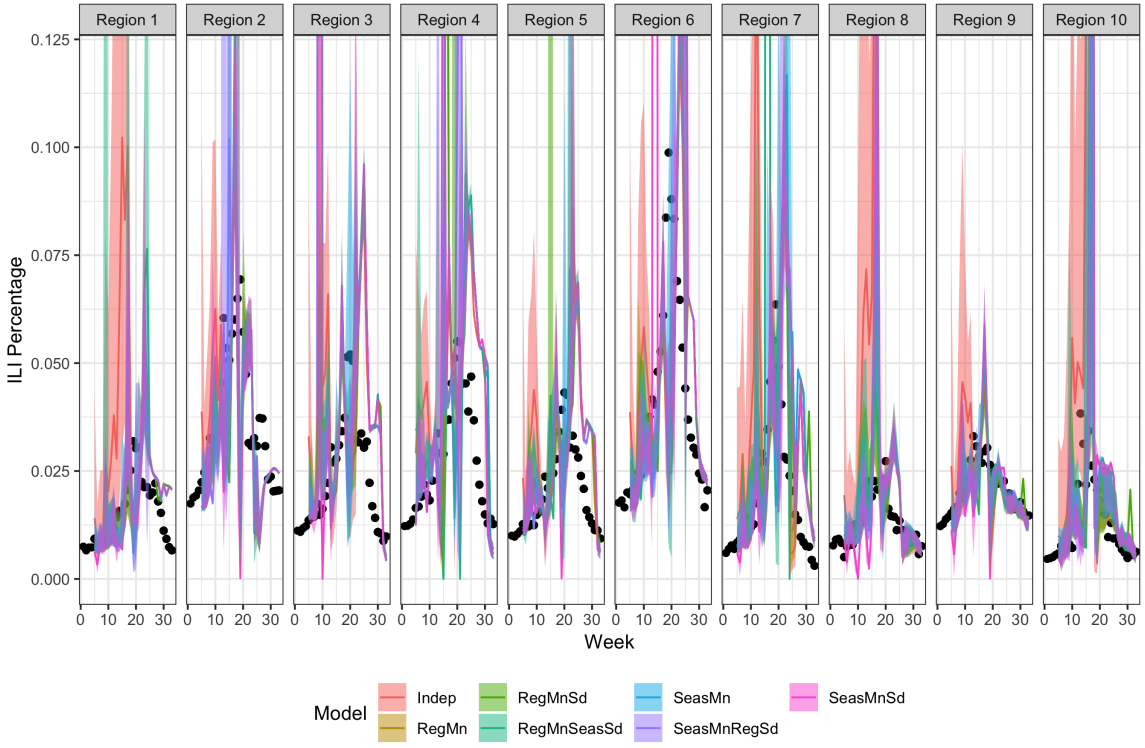

Figure .9 4 week ahead forecasts and their 95% credible intervals for all regions in the $2016-2017$ influenza season.

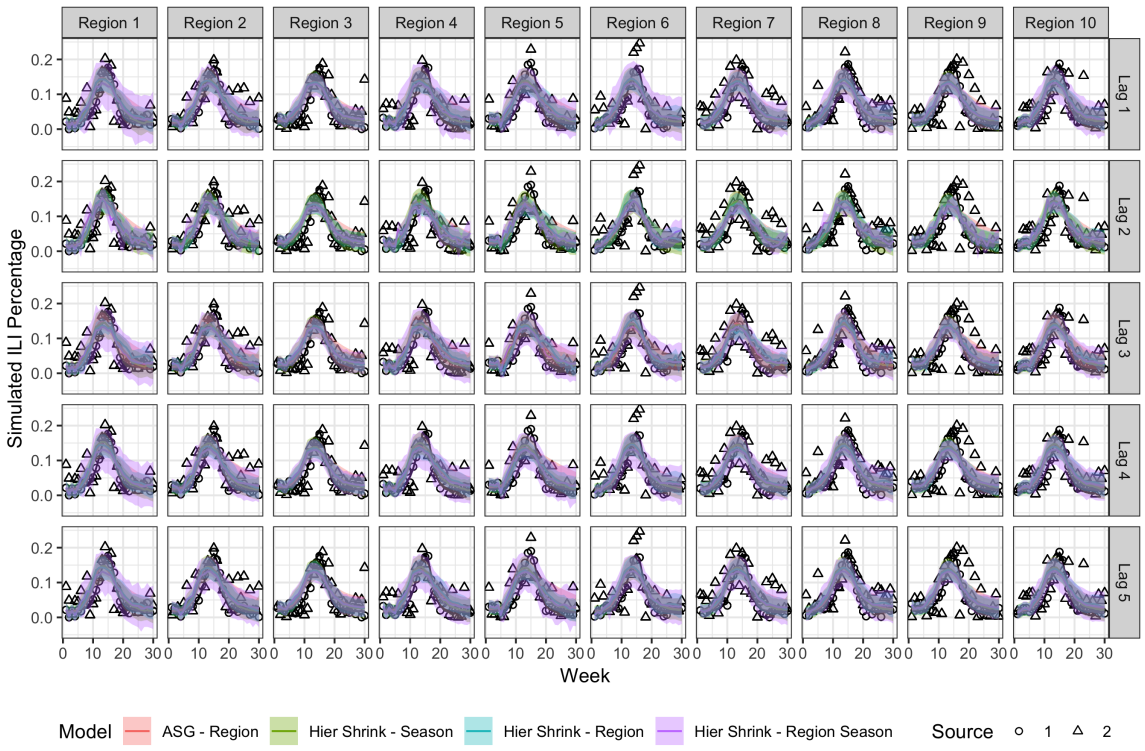

Figure .10 Posterior mean fit on the simulated data when including 5 weeks of the forecasted season in the forecast. The columns are faceted by region and the rows are faceted by lag.

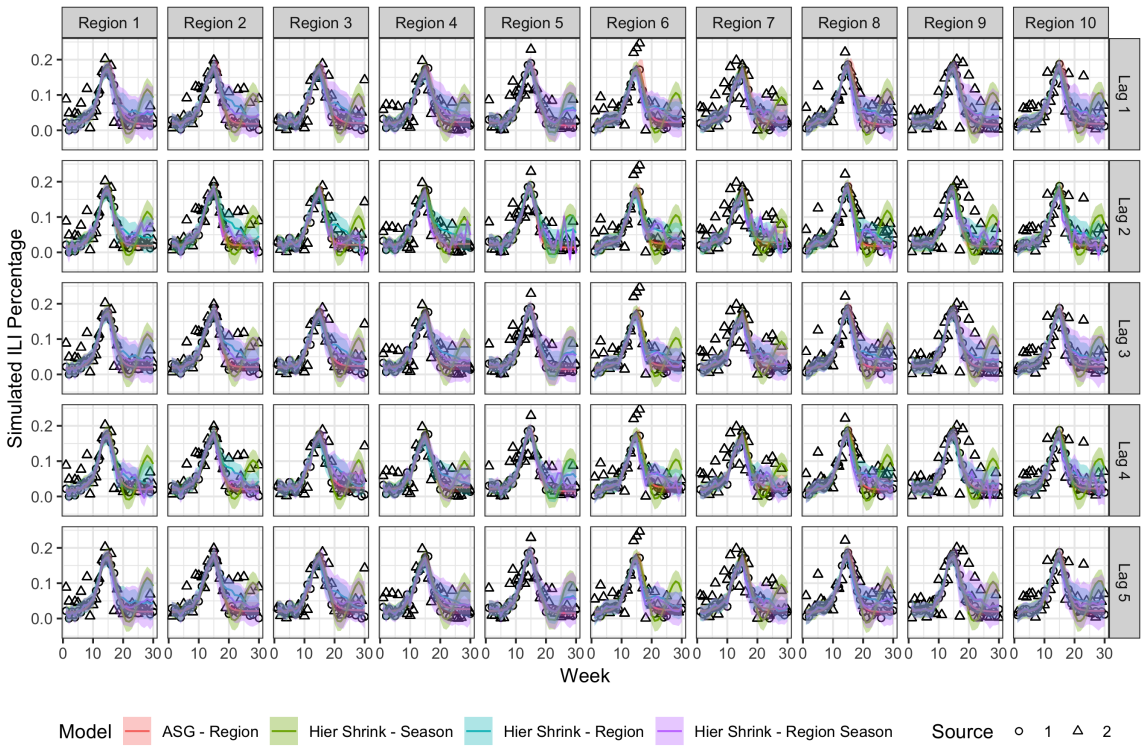

Figure .11 Posterior mean fit on the simulated data when including 15 weeks of the forecasted season in the forecast. The columns are faceted by region and the rows are faceted by lag.

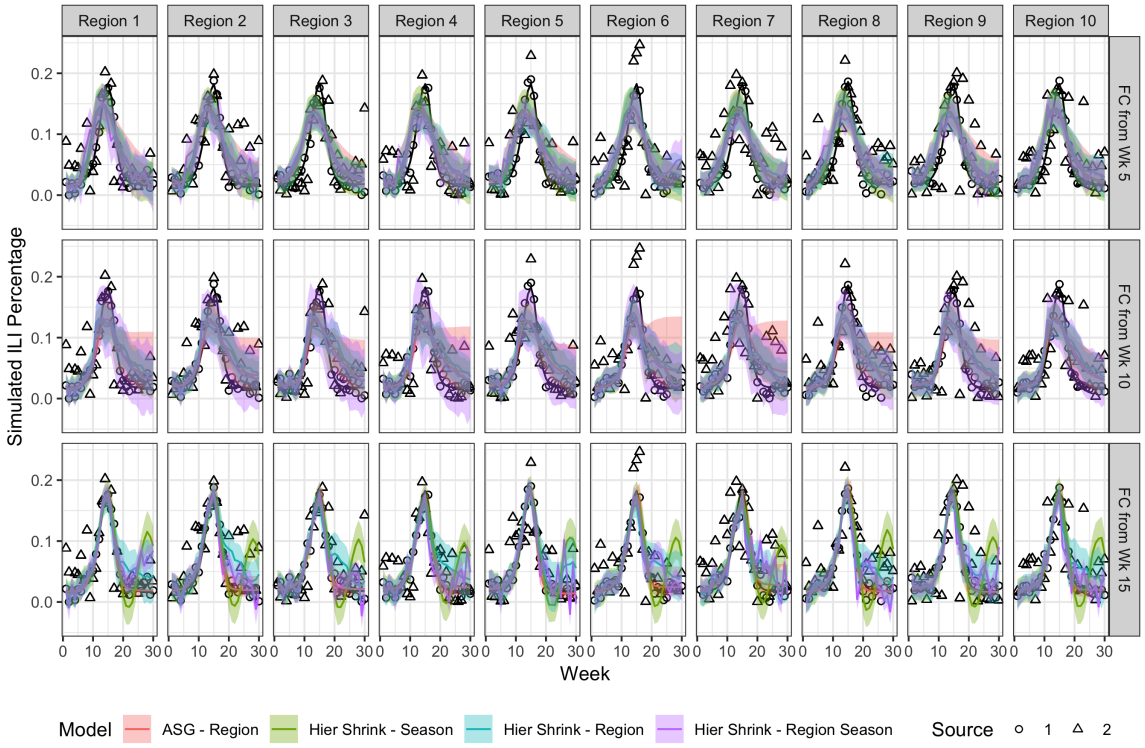

Figure .12 Posterior mean fit on the simulated data when including 2 weeks of lag in the forecast. The columns are faceted by region and the rows are faceted by number of weeks included in the forecast.

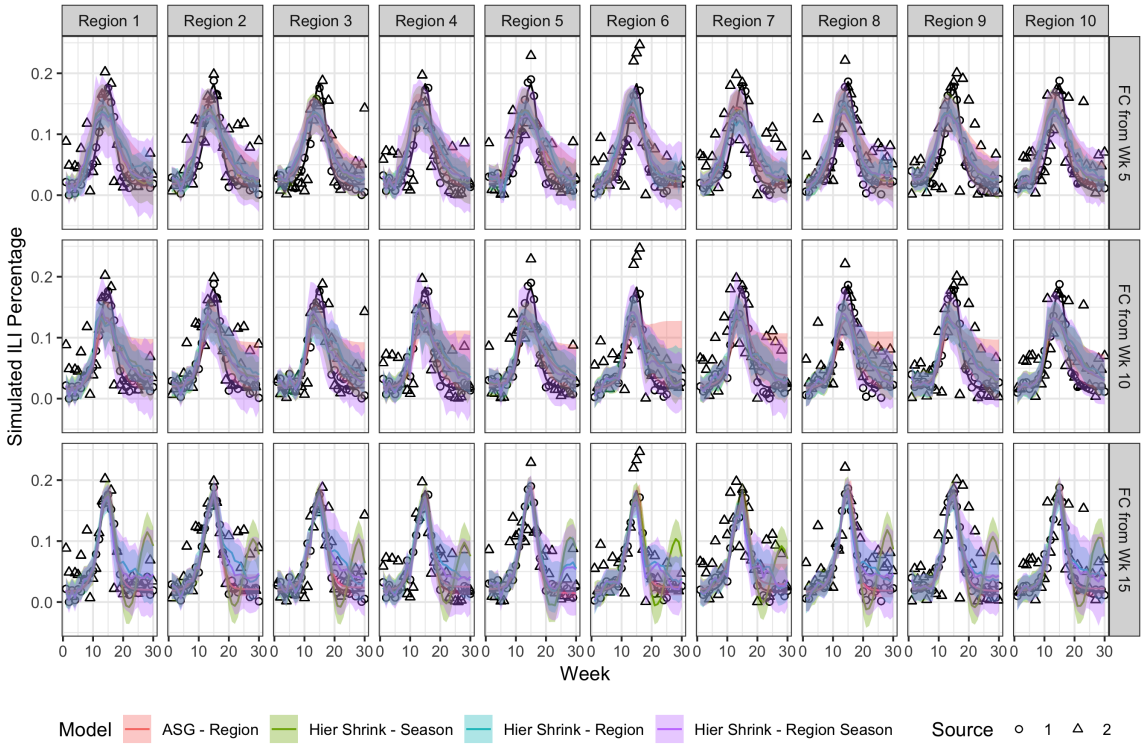

Figure .13 Posterior mean fit on the simulated data when including 3 weeks of lag in the forecast. The columns are faceted by region and the rows are faceted by number of weeks included in the forecast.

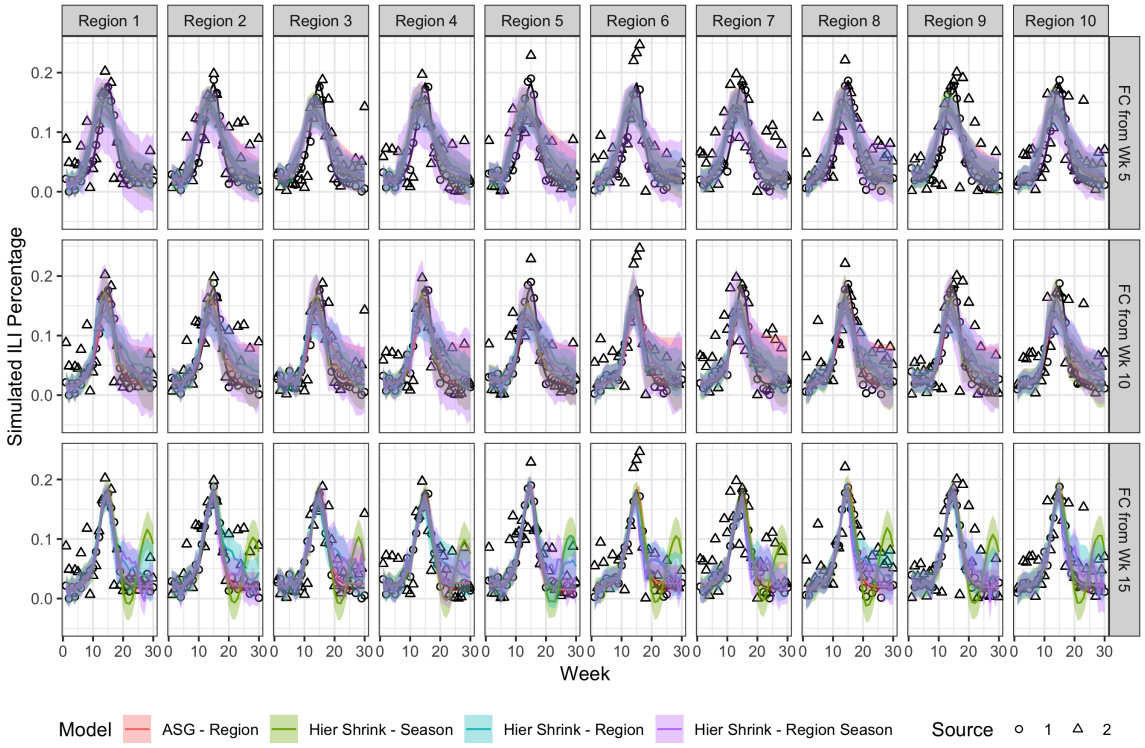

Figure .14 Posterior mean fit on the simulated data when including 4 weeks of lag in the forecast. The columns are faceted by region and the rows are faceted by number of weeks included in the forecast.

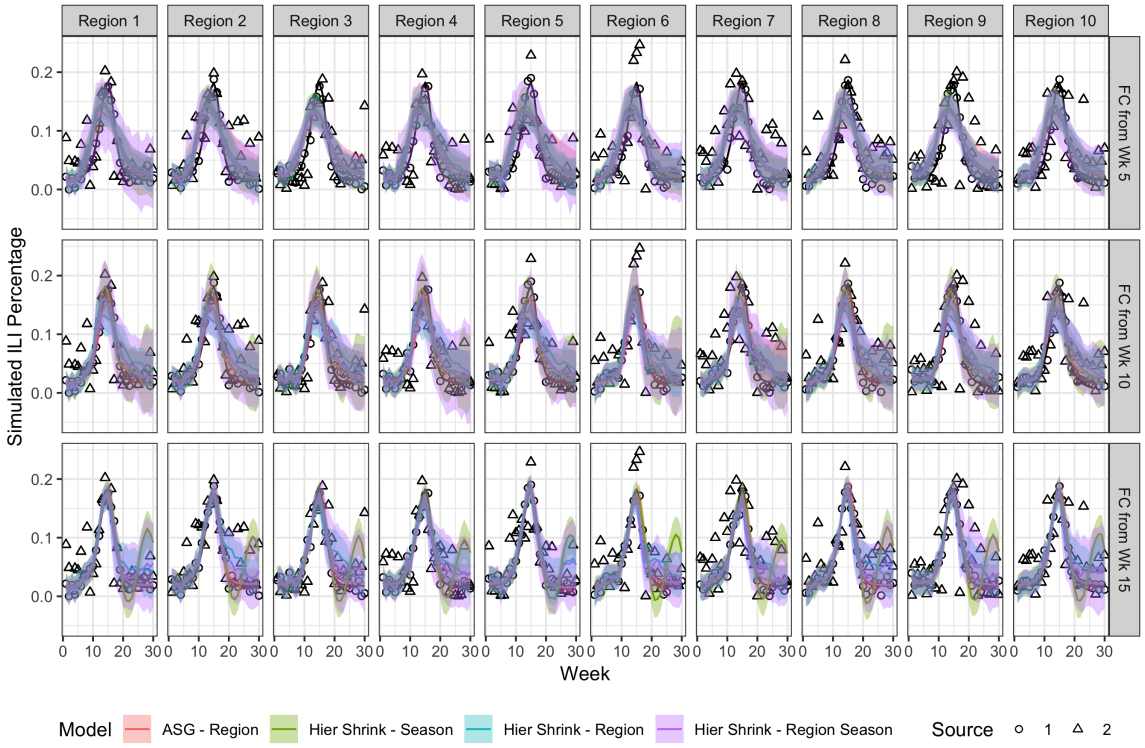

Figure .15 Posterior mean fit on the simulated data when including 5 weeks of lag in the forecast. The columns are faceted by region and the rows are faceted by number of weeks included in the forecast.